Analysis of National Weather Service Stage Forecast Errors

GRACE ZALENSKI, WITOLD F. KRAJEWSKI, AND FELIPE QUINTERO

Iowa Flood Center, University of Iowa, Iowa City, Iowa

PEDRO RESTREPO AND STEVE BUAN

North Central River Forecast Center, National Weather Service, Chanhassen, Minnesota

(Manuscript received 16 December 2016, in final form 24 April 2017)

ABSTRACT

This paper explores the skill of river stage forecasts produced by the National Weather Service (NWS). Despite the importance of the verification process in establishing a reference that allows advancement in river forecast technology, there is relatively little literature on this topic. This study aims to contribute to this subject. The study analyzed the North Central River Forecast Center's river stage forecasts for 51 gauges in eastern and central Iowa between 1999 and 2014. The authors explored forecast skill dependence characteristics such as upstream area, water travel time, and the number of gauges located upstream of each forecasting point. They also assessed the influence of rainfall uncertainty on stage error by examining the relationship between the forecast skill and its antecedent 24-h observed rainfall. The results show that when using persistence as a reference for comparison with NWS actual forecasts, the NWS forecasts are better for predictions below and above flood stage. The difference in root-mean-square error (RMSE) between the actual and persistence forecasts ranges between 0.04 and 1.24 ft, and it increases with lead time. Locations with fewer upstream gauges exhibit greater variation in forecast skill than locations that are well gauged, especially at high flood levels. Strong predictive relationships between the physical characteristics of a basin (travel time, upstream drainage area), rainfall quantities, and forecast skill have not been identified.

1. Introduction

The National Weather Service (NWS) has the mandate of providing streamflow forecast services for the United States. To meet its function, it relies on observations from the U.S. Geological Survey (USGS), the U.S. Army Corps of Engineers, and a number of federal, state, and tribal partners that supply streamflow, snow, temperature, and other variables used in the forecast process. Hydrologic forecasts produced by the NWS in real time, specifically river streamflow forecasts, are used by decision-makers and the general public for a variety of purposes, from routine flow control operations to preparing and executing emergency response strategies to extreme hydrologic events. Reliable forecasts, specifically those of river stage, are of particular importance during these extreme events.

While the literature is abundant in description and performance evaluation of hydrologic rainfall–runoff models, surprisingly little has been documented about the real-time forecasting skill of the models used by the NWS. Notable exceptions are the works by Welles and Sorooshian (2009) and Welles et al. (2007). Welles et al. (2007) published a verification study using 10 years' worth of data from four forecast sites in Oklahoma, and 20 years of data from 11 sites along the Missouri River. They found that below flood stage, NWS forecasts demonstrate some skill for 1-, 2-, and 3-day forecasts, but above flood stage, only the 1-day forecasts show skill. After these studies, there have not been any other comprehensive papers evaluating the skill of NWS forecasts despite the fact that the number of years of data available for analysis and the number of stream gauges to evaluate have increased with time.

In this paper we document the results of a study exploring forecasting error of river stage forecasts in Iowa for the years 1999–2014. During this time, Iowa experienced numerous floods all across the state. Figure 1 shows the number of flood-related presidential disaster declarations issued in the 99 counties between 1964 and 2010. There are indications that the wet trend in Iowa

Corresponding author: Felipe Quintero, felipe-quintero@uiowa.edu

DOI: 10.1175/WAF-D-16-0219.1

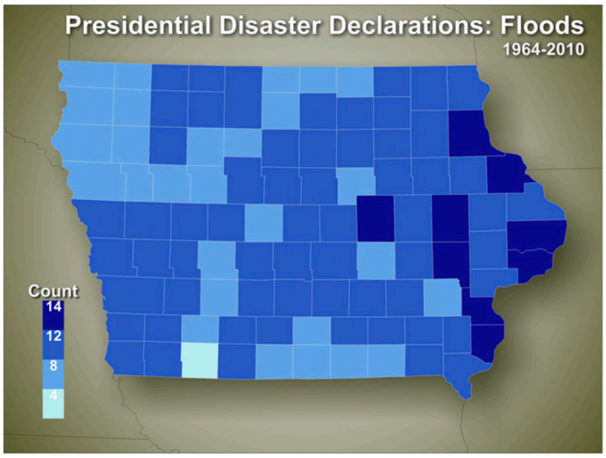

FIG. 1. Number of federally declared IA flood disasters by county.

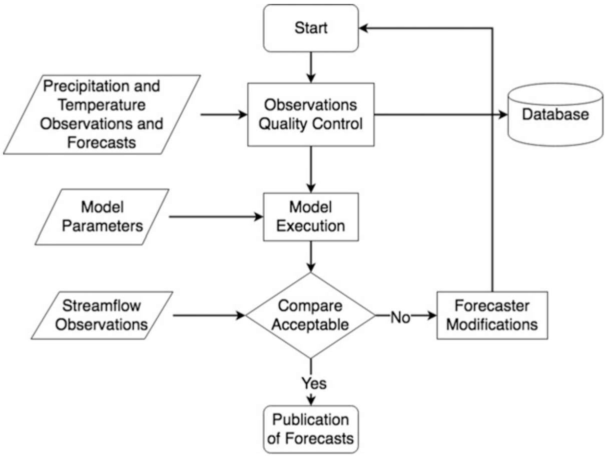

FIG. 2. NWS forecast flowchart at the NCRFC.

will continue, and the state should expect more floods to come (e.g., Villarini et al. 2011a; Mallakpour and Villarini 2015; Villarini et al. 2011b.) Therefore, it is of interest to examine our ability to forecast river stage at flood levels within the region.

We are also motivated by the need to establish a reference level for future advances in forecasting technology. The Iowa Flood Center, established by the state in 2009 has developed and operates a detailed, state-wide streamflow forecasting model that updates its forecasts every hour for thousands of locations across Iowa (Krajewski et al. 2016). The recently established National Water Center (NWC) in Tuscaloosa, Alabama, is developing a new method of flow prediction, which may have major impacts on the process of streamflow forecasting (NOAA 2016). As new models and methods come on line and their performance is assessed, an obvious question arises: Do they perform better than the current practice? But such reference standards are difficult to find, and we would like this study to at least partially fill the gap.

Forecast verification—assessing the reliability and accuracy of hydrologic forecasts—is an important component of the forecasting process. In addition to giving a measure of confidence to those utilizing hydrologic forecasts, verification provides necessary information and feedback for forecasters and model developers. However, there is not currently a standardized, well-defined method of river stage forecast verification. Single statistics such as RMSE, which are most commonly used for verification, are not sufficient to completely characterize forecast skill and tend to emphasize the calibration of high flows. In this study, we are interested in both assessing the overall predictive power of hydrologic forecasts and comparing the quality of forecasts issued for different basins, which may highlight areas of interest for further studies.

Comparing forecast error across basins is a major challenge in hydrologic forecast verification. While flood stage is a useful and practical measurement, it is not itself a conservative quantity, so it may be more difficult to identify relationships between stage forecast error and the physical characteristics of a basin, such as drainage area. Still, stage is the quantity that the general public and emergency responders, who often lack hydrologic background and training, readily understand. Therefore, we decided to analyze the past forecast data in terms of errors of the stage. We attempted to perform several analyses to explain the results and gain insights into what rules good versus not-so-good performance. We elaborate later in the paper.

2. Forecasting methods and procedures used at North Central River Forecast Center

The heart of the forecast process is the Community Hydrologic Prediction System (CHPS), which is supported by the infrastructure of the Flood Early Warning System (FEWS), developed by Deltares in the Netherlands (http://oss.deltares.nl/web/delft-fews/about). FEWS is widely used around the world by flow forecasting agencies.

The forecasting duties for major rivers are distributed among 13 NWS River Forecast Centers (RFCs) across the United States. Flash floods, on the other hand, are the responsibility of the 122 NWS Weather Forecast Offices, with guidance and support provided by the RFCs. This paper is focused on forecasts and forecasting procedures at the North Central River Forecast Center (NCRFC) in Chanhassen, Minnesota.

TABLE 1. Description of the 51 NWS forecast locations used in the study, and the time period where forecasts were issued where available. Benchmark locations are marked in boldface.

	NWS ID	River	Town	Area (mi^2)	Initial forecast	Last forecast
1	DCHI4	Upper Iowa	Dorchester	769	16 May 1999	10 Sep 2014
2	DEHI4	Upper Iowa	Decorah	491	19 Jul 1999	19 Jun 2014
3	GRBI4	Turkey	Garber	1557	16 May 1999	3 Jul 2014
4	EKDI4	Turkey	Elkader	911	8 Jul 2003	10 Sep 2014
5	MAQI4	Maquoketa	Maquoketa	1550	22 Apr 1999	10 Sep 2014
6	MCHI4	Maquoketa	Manchester	282	16 Jul 2004	1 Jul 2014
7	**DEWI4**	Wapsipinicon	De Witt	2287	7 Sep 1998	17 Sep 2014
8	ANSI4	Wapsipinicon	Anamosa	1581	16 Jul 2004	17 Sep 2014
9	IDPI4	Wapsipinicon	Independence	1052	16 May 1999	5 Jul 2014
10	CNEI4	Cedar	Conesville	7753	8 Apr 1999	12 Sep 2014
11	**CIDI4**	Cedar	Cedar Rapids	6492	16 May 1999	17 Sep 2014
12	**ALOI4**	Cedar	Waterloo	5133	10 Feb 1999	9 Jul 2014
13	NHRI4	Cedar	New Hartford	349	10 Feb 1999	4 Jul 2014
14	FNHI4	Cedar	Finchford	852	8 Apr 1999	8 Jul 2014
15	JANI4	Cedar	Janesville	1676	22 Apr 1999	30 Jun 2014
16	SHRI4	Cedar	Shell Rock	1711	6 Apr 1999	4 Jul 2014
17	WVLI4	Cedar	Waverly	1563	26 Apr 2009	17 Jul 2011
18	CCYI4	Cedar	Charles City	1077	8 Apr 1999	23 Jun 2014
19	MCWI4	Cedar	Mason City	484	10 Feb 1999	4 Jul 2014
20	**WAPI4**	Iowa	Wapello	12483	1 Apr 1999	17 Sep 2014
21	**LNTI4**	Iowa	Lone Tree	4291	23 Apr 1999	13 Sep 2014
22	KALI4	Iowa	Kalona	576	23 Apr 1999	13 Sep 2014
23	**IOWI4**	Iowa	Iowa City	3267	22 Mar 2002	17 Sep 2014
24	MROI4	Iowa	Marengo	2795	8 Apr 1999	11 Sep 2014
25	**MIWI4**	Iowa	Marshalltown	1534	1 Apr 1999	17 Sep 2014
26	ROWI4	Iowa	Rowan	426	8 Mar 2010	10 Sep 2014
27	AGSI4	Skunk	Augusta	4333	21 Apr 1999	14 Sep 2014
28	SIGI4	Skunk	Sigourney	735	21 Apr 1999	14 Sep 2014
29	OOAI4	Skunk	Oskaloosa	1657	8 Apr 1999	11 Sep 2014
30	AESI4	Skunk	Ames	563	8 Apr 1999	10 Sep 2014
31	AMWI4	Skunk	Ames	213	8 Apr 1999	10 Sep 2014
32	AMEI4	Skunk	Ames	314	22 Apr 1999	3 Jul 2014
33	**KEQI4**	Des Moines	Keosauqua	14083	26 Feb 1997	17 Sep 2014
34	**OTMI4**	Des Moines	Ottumwa	13421	26 Feb 1997	16 Sep 2014
35	BSSI4	Des Moines	Bussey	374	26 Feb 1997	11 Sep 2014
36	TRCI4	Des Moines	Tracy	12526	26 Feb 1997	30 Jul 2008
37	AKWI4	Des Moines	Ackworth	461	26 Feb 1997	11 Sep 2014
38	IDNI4	Des Moines	Indianola	499	26 Feb 1997	12 Sep 2014
39	NRWI4	Des Moines	Norwalk	350	26 Feb 1997	14 Sep 2014
40	**DESI4**	Des Moines	Des Moines	9926	26 Feb 1997	11 Sep 2014
41	**DMOI4**	Des Moines	Des Moines	6305	26 Feb 1997	3 Jul 2014
42	GRMI4	Des Moines	Grimes	371	26 Feb 1997	10 Sep 2014
43	STRI4	Des Moines	Stratford	5500	26 Feb 1997	7 Jul 2014
44	WBCI4	Des Moines	Webster City	840	26 Feb 1997	1 Jul 2014
45	HBTI4	Des Moines	Humboldt	2286	26 Feb 1997	9 Jul 2014
46	DAKI4	Des Moines	Dakota City	1304	26 Feb 1997	22 Jun 2014
47	**DEMI4**	Raccoon	Des Moines	3612	26 Feb 1997	11 Sep 2014
48	DMWI4	Raccoon	Des Moines	3517	14 Dec 2004	10 Sep 2014
49	**VNMI4**	Raccoon	Van Meter	3436	26 Feb 1997	10 Sep 2014
50	REDI4	Raccoon	Redfield	972	26 Feb 1997	27 Aug 2014
51	EFWI4	Raccoon	Jefferson	1658	26 Feb 1997	30 Jun 2014

Forecasting process

Figure 2 shows a schematic of the forecasting process for deterministic forecasts, which is the topic of this paper. The first step in the forecasting work flow is the quality control primarily of precipitation observations, although temperature observations may also require a quality checkup. Precipitation and temperature observations and forecasts are delivered to each RFC via the internal NWS-wide area network. Currently, the

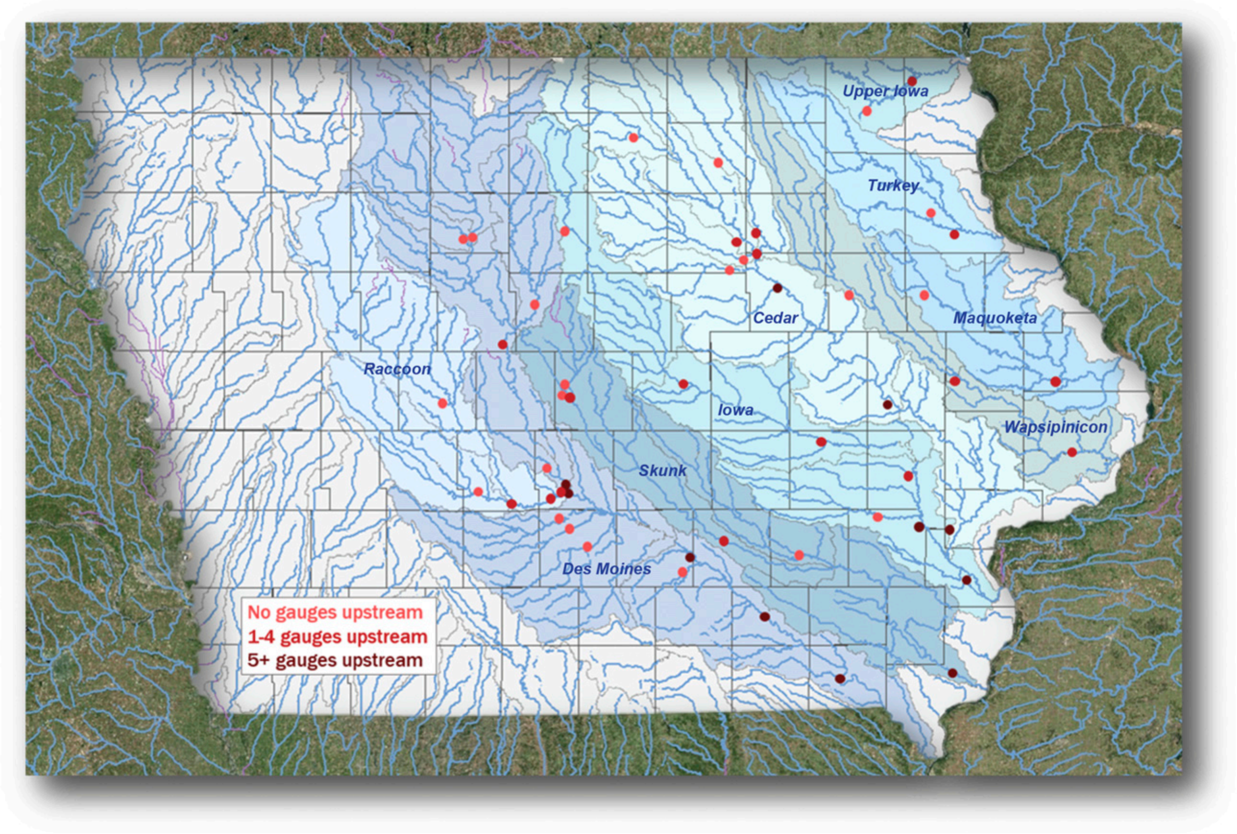

FIG. 3. NWS forecast locations used for this study. Sites with no upstream gauges are marked in light red, sites with between one and four upstream gauges are in red, and sites with five or more upstream gauges are in dark red. Watersheds for major IA rivers (from NE to SW: Upper Iowa, Turkey, Maquoketa, Wapsipinicon, Cedar, Iowa, Skunk, Des Moines, and Raccoon) are shaded in blue.

operational hydrologic forecasting models only use precipitation and temperature observations and forecasts, although the human forecasters do examine additional variables such as soil moisture and snow conditions.

Quality control of observations is performed by the Hydrometeorological Analysis and Support (HAS) forecaster, who uses tools to provide either a point or areal correction of observations. Point observations are obtained from ground-based rain gauges and climate reference stations. Spatial observations of precipitation are obtained primarily from a network of 150 NEXRAD radars, with additional information from the network of Environment and Climate Change Canada (formerly known as Environment Canada) radars located near the U.S. border. In addition, satellite observations from National Oceanic and Atmospheric Administration's (NOAA) geostationary satellites and now the Global Precipitation Measurement Mission (GPM) of NASA and the Japan Aerospace Exploration Agency (JAXA) provide additional precipitation information in areas of sparse coverage. Quality-controlled observations are communicated to the National Centers for Environmental Prediction (NCEP) and to NOAA's National Centers for Environmental Information (NCEI) for archival and distribution.

Precipitation and temperature forecasts are received from NCEP. Although those forecasts have already been tweaked by the NCEP forecasters (Novak et al. 2014), there may be a need for additional adjustment by the local forecasters. This is done by the HAS forecaster, the hydrologic forecaster, and, frequently, in collaboration with the weather forecast offices in the watershed.

Once the process of quality controlling the observations is complete, the actual forecast runs can proceed. The NCRFC uses a number of models in its operational hydrologic forecasting process. The most important models are the Snow Accumulation and Ablation Model (SNOW-17; Anderson 1976), for snow accumulation and melting; the Sacramento Soil Moisture Accounting model (SAC-SMA; Burnash 1995) for the accounting of soil moisture processes, surface and direct runoff, infiltration, evapotranspiration,

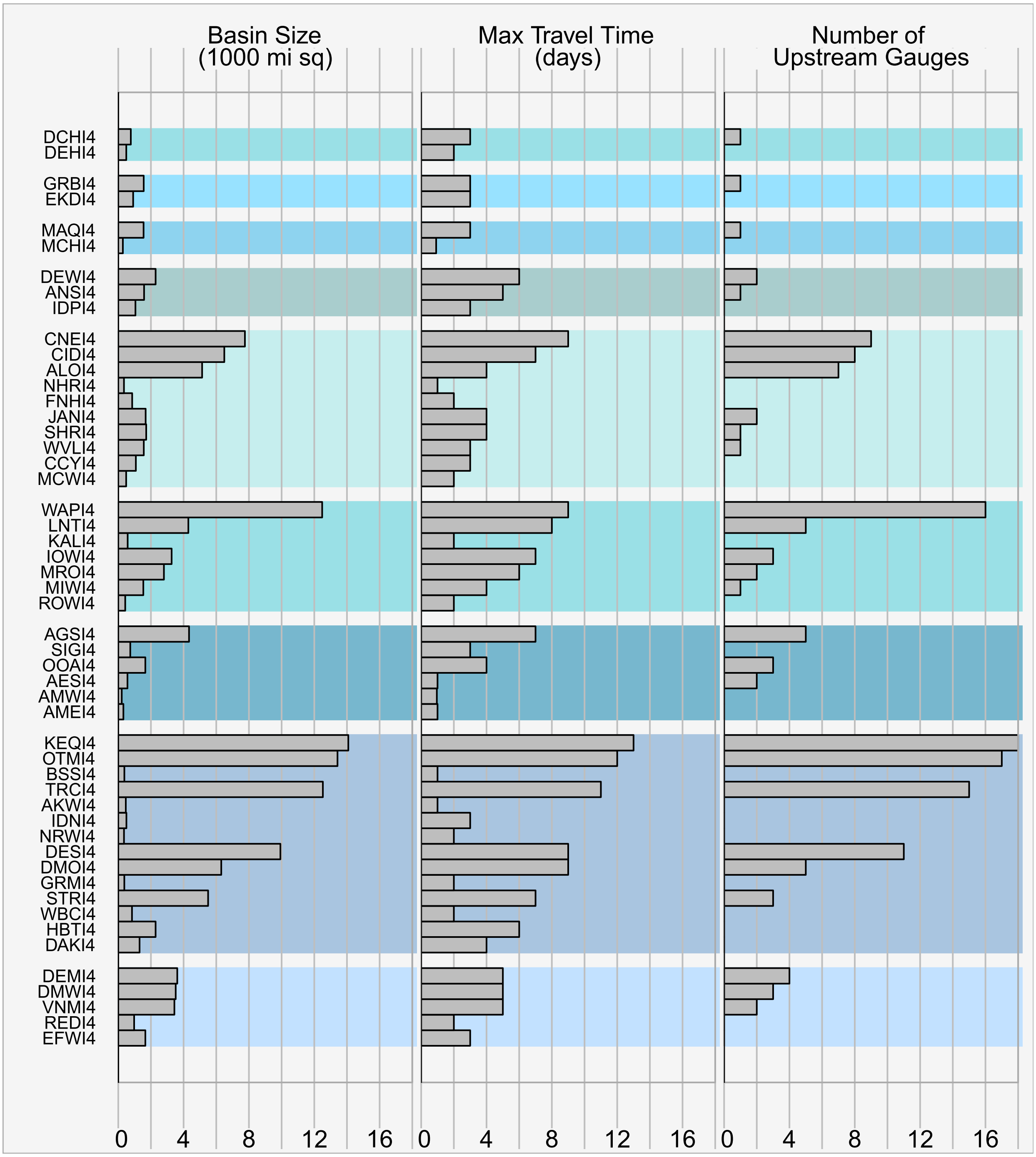

FIG. 4. Characteristics of forecast points: (left) drainage area upstream of each site, (center) estimated travel time along the maximum stream distance through the drainage area upstream of each site, and (right) number of NWS gauges used in this study upstream of each site. Gauges are grouped according to the major IA river each drains into.

and groundwater contribution to streamflow; the Hydrologic Engineering Center's River Analysis System (HEC-RAS; Brunner 2010) for hydraulic routing; several reservoir routing models, such as the Joint Reservoir Regulation Operation (RES-J) and the HEC Reservoir System Simulation (HEC-ResSim) model (Klipsch and Hurst 2007); and, finally, a number of other hydrologic routing and time series manipulation routines, including the unit hydrograph model.

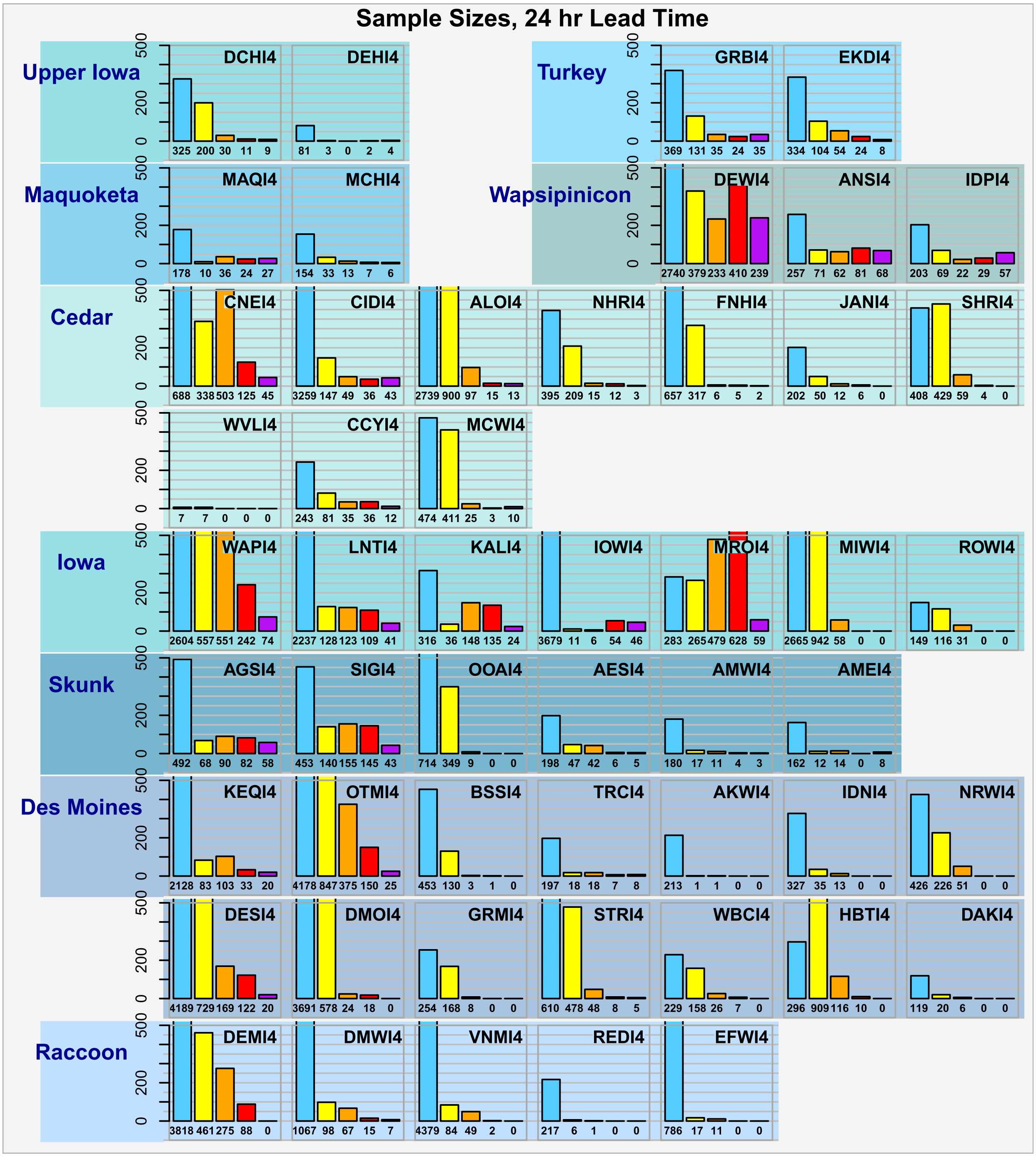

FIG. 5. Sample sizes from each site, stratified by flood levels as defined by the NWS. Blue for below flood level, yellow for above action level, orange for above flood stage, red for above moderate flood, and purple for above major flood. We count only forecast–observation pairs from the 1-day forecasts, so total sample size equals the number of forecasts for the sample period.

River basins are divided into subbasins, and executed by the suite of uniquely calibrated models under CHPS control from upstream to downstream, to allow flow accumulation and routing from upstream subbasins to be added to the local contributions of downstream basins. The suite of models is executed in sequence, starting with SNOW-17, followed by SAC-SMA, the unit hydrograph model reservoir routing models, and, where required, HEC-RAS or, alternatively, any of the hydrologic routing methods. The parameter set for each of the

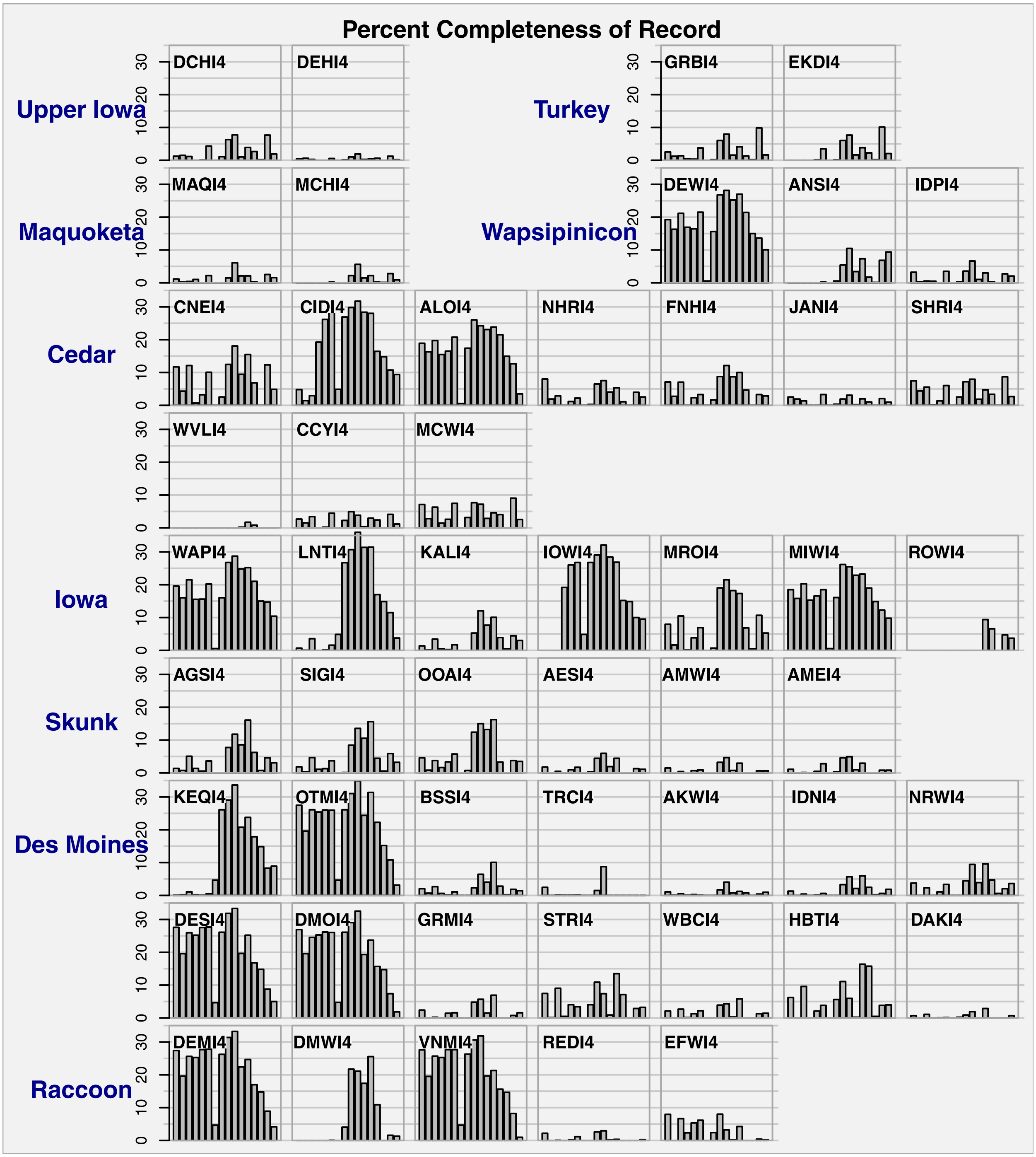

FIG. 6. Percent completeness of the data record at each site, by year. Sample period is 16 yr, 1999–2014. We define 100% completeness as a forecast issued every 6 h, every day of the year. Percent completeness is the percentage of forecasts issued out of the maximum possible number.

models and for each subbasin is used for the model execution, and the results are made available to the forecaster both in text and graphical formats.

The next step is a very important step in the operational hydrologic forecasting process of the NWS. The forecaster inspects the numerical forecast and compares its performance with past observations. Based on the comparison, the forecaster may decide to perform some modifications. Those may be as simple as the use of a blending model that adjusts the hydrologic-model-simulated flows in the recent

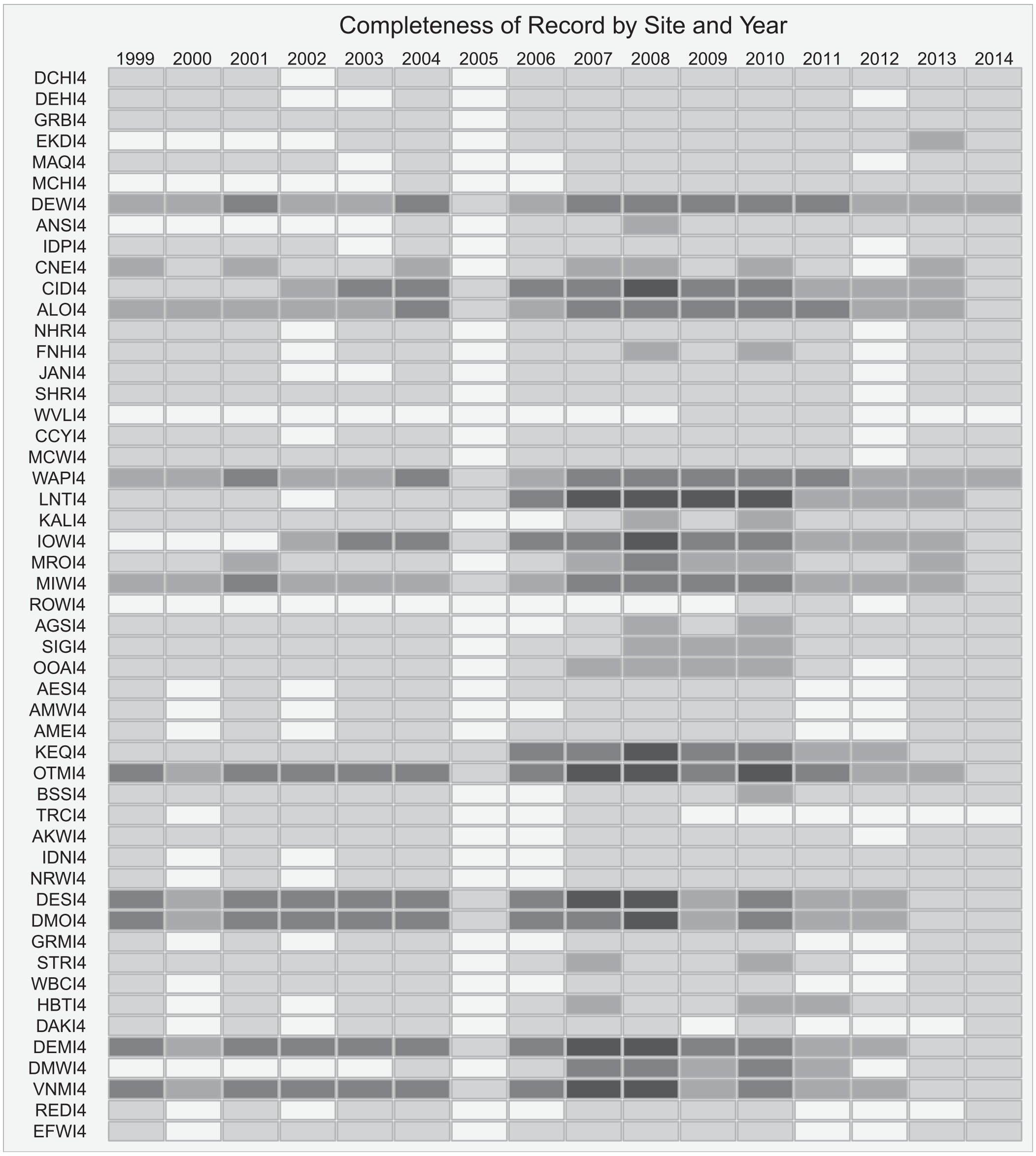

FIG. 7. Percent completeness of the data record by site and year. Boxes with no color represent years with no data reported. Darker shades correspond to a higher degree of completeness. Categories are <10% complete, 10%–20% complete, 20%–30% complete, and 30%–40% complete.

past time steps to match the streamflow observations, such that for the forecasts, the model will slowly move toward the original forecasted value.

Other types of modifications (MODS) can be performed. For instance, the forecaster observes that a storm centroid is focused either at the head or the mouth of the watershed. Then, it is possible to modify the unit hydrograph to reflect the shape of the basin response. Also, the forecaster can modify model parameters, or even observed precipitation estimates. The new values are put back into the data, or parameter input files, and the model is executed again. Once the forecaster judges that the comparison is acceptable, the forecast is made official and is distributed to the corresponding weather forecast offices for publication online.

3. Data summary

a. *Forecasting locations*

The data for this study come from 51 National Weather Service forecast locations in east and central Iowa under the responsibility of the NCRFC (Table 1; Fig. 3). The dataset consists of river stage forecasts issued at these sites for the years 1999–2014. The forecast

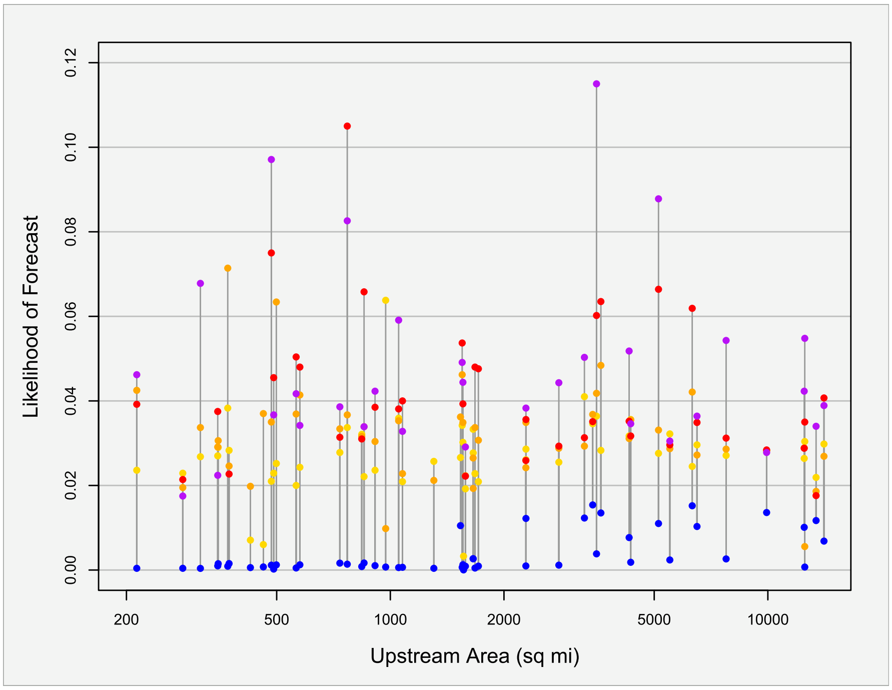

FIG. 8. Likelihood of forecast by flood stage, plotted against the upstream drainage area. Likelihood is the fraction of particular observations at a particular flood level for which there is a corresponding forecast. Stage observations are recorded every 15 min, and forecasts are issued at most every 6 h, so we expect small values. Likelihood of forecast for observations below flood level in blue, at action level in yellow, at flood stage in orange, at moderate flood stage in red, and at major flood stage in purple.

horizon changed during this period. Forecasts from 1999 through April 2004 were issued with a 120-h time horizon, and from May 2004 to 2014 the forecast horizon was extended to 168 h. The river stage observations from USGS gauges at the forecast points are available for the same period. Forecast points used for this study are shown in Fig. 3. In Fig. 3 we distinguish between forecast locations with and without other forecast points located in their upstream basins.

There are two distinct categories of forecast point included in this study. Thirteen of the 51 locations are NWS benchmark locations (see Table 1). The procedure for these locations is to issue at least one forecast per day between 1 April and 31 October, regardless of current conditions. This ensures that a more complete and geographically distributed dataset is available for analysis, even in the absence of hydrologic events. Outside of that date range, and for all other forecast locations, a forecast would only be issued if the observed or predicted stage was above "action stage."

The benchmark forecast locations in Iowa are as follows: De Witt (DEWI4), Cedar Rapids (CIDI4), Waterloo (ALOI4), Wapello (WAPI4), Lone Tree (LNTI4), Iowa City (IOWI4), Marshalltown (MIWI4), Keosauqua (KEQI4), Ottumwa (OTMI4), Des Moines (DESI4, DMOI4, DEMI4), and Van Meter (VNMI4).

Figure 4 shows the upstream drainage area, approximate water travel time, and number of gauges located upstream of each forecast point. The drainage areas of these sites range from 213 to 14 083 mi^2 (552–36 475 km^2). Travel times range from less than a day to 13 days, and the number of upstream gauges ranges from 0 to 18. These three point characteristics are closely related to each other, and each characteristic may influence forecast skill at a given site. For instance, sites with larger drainage areas may be more predictable in their response to rainfall, but forecasting efforts must take into account the complexity resulting from the confluence of multiple tributaries within the basin. Forecast points with longer maximum travel time benefit from advance knowledge of rainfall from farther upstream. We also hypothesize that forecast skill at a site is related to the presence of gauges upstream of that location, as a result of the stage and discharge information provided by upstream gauges.

NWS forecasts consist of stage projections at 6-h intervals from the time of issue. Forecasts generally extend 5–7 days into the future and may be issued up to four times a day. Stage observations are recorded every

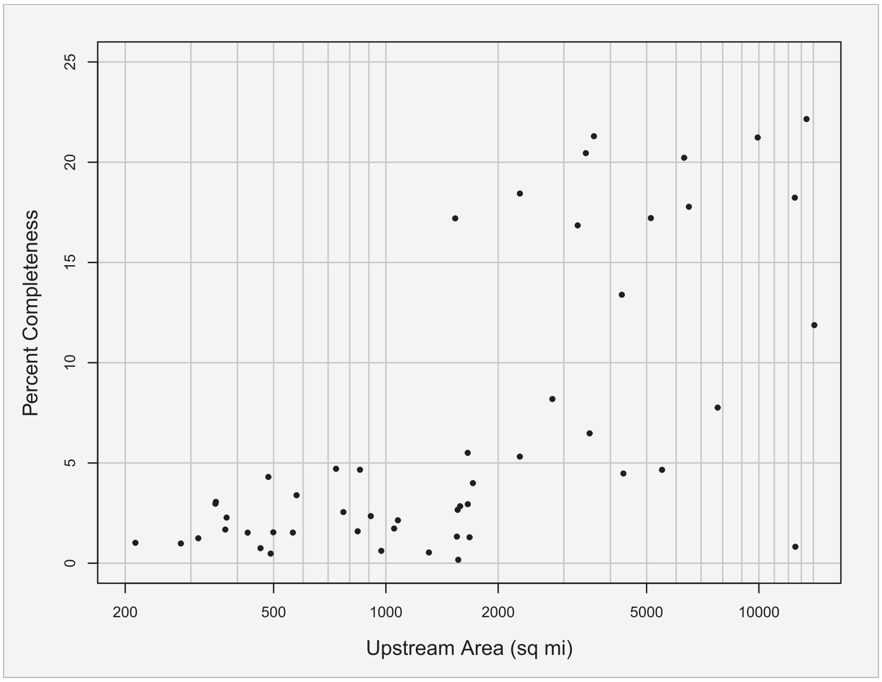

FIG. 9. Scatterplot displaying percent completeness of the record for each site against the site's upstream drainage area.

15 min. For this study, stage was back calculated from discharge data using USGS rating curves because of the difficulty in obtaining historically observed stage data. To assess the NWS forecasts, we build a relational database of forecast–observation pairs using the stage observation closest in time to the projection. The pairs were stratified by lead time and sorted by the observed value into five flood stage categories (below, action, flood stage, moderate flood, and major flood). The categories are NWS-defined flood stage levels unique to each site. We can then calculate and compare the error distributions at every flood stage with respect to important site characteristics. The use of a relational database for managing the data provides flexibility on the

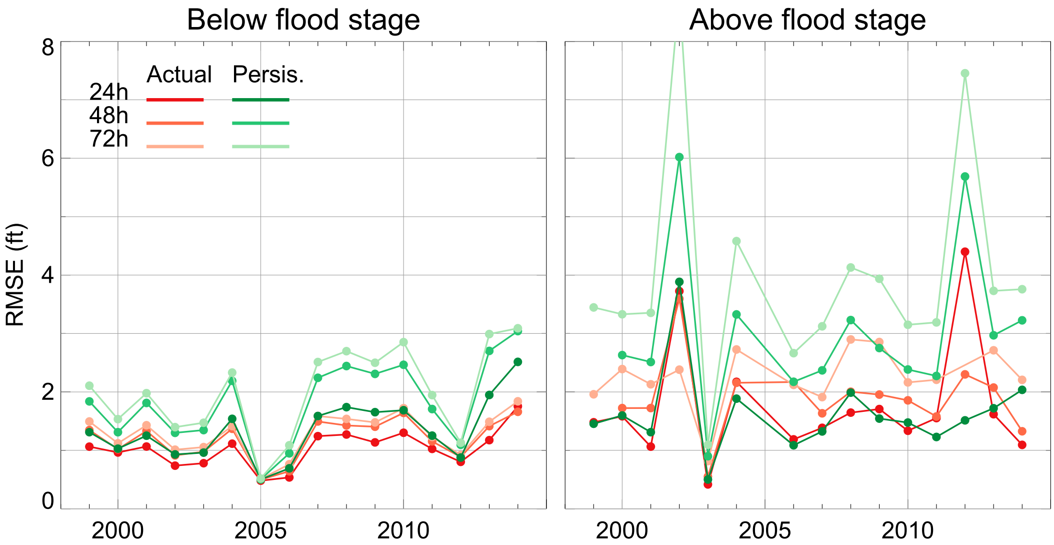

FIG. 10. Median RMSE of actual and persistence forecasts where the stage at time of issue was below or above flood level. Median RMSE is obtained for each year, using all forecasts issued at all forecast points for that year, at 1-, 2-, and 3-day lead times.

TABLE 2. Median RMSE (in ft) of actual and persistence forecasts for different lead times.

Forecast	Lead time (h)	Below flood stage	Above flood stage
Actual	24	1.06	1.55
Persistence		1.31	1.51
Actual	48	1.34	1.95
Persistence		1.83	2.75
Actual	72	1.42	2.20
Persistence		2.10	3.44

conditioning for the analyses (e.g., flood stage categories, basin characteristics, year when the forecast was issued, among others).

b. *Rainfall data*

NWS streamflow forecasts also take into account 24-h forecasts of expected precipitation. Forecasted precipitation is updated with actual precipitation amounts as time progresses and new forecasts are issued.

To assess the influence of rainfall uncertainty on stage error, we used Stage IV, the national rainfall product available for the period 2002–14. The rainfall product is derived from the hourly multisensor (radar and rain gauges) precipitation estimates that are produced and distributed by NCEP (Lin and Mitchell 2005). The hourly rainfall estimates were spatially averaged over the drainage area for each forecast point. We discuss the use of the radar rainfall data in section 5.

4. Forecast sampling

It is important to note that sample size—the number of forecast–observation pairs—varies significantly between flood categories and between sites (Fig. 5). Because of the nature of flooding, a majority of observations at almost every site are below flood stage. The exception is Marengo (MROI4), which is directly upstream of Coralville Lake, a major reservoir for the area. The large number of observations recorded at flood levels may reflect the need to closely monitor and manage reservoir intake under flood conditions. Controlled flow from the reservoir directly upstream (12 km) also accounts for the unusually small number of observations at low flood levels at the Iowa City gauge (IOWI4). We also see that the sample size varies considerably from site to site. In general, forecast points near the outlet of major rivers have larger sample sizes. Sites located near cities with large populations (Des Moines, Iowa City, Waterloo) are also likely to have a large number of observations. Many sites, in particular those with small drainage areas, have relatively few observations at any river stage. In these cases there may be no record of any forecast–observation pairs corresponding to the higher flood levels. However, there exists the possibility that a flood event occurred, but no forecast was issued to capture it. This is described by Welles et al. (2007) as the "no forecast problem." We found that on average over all the sites, there were observed river stages above flood stage category with no available forecasts during 3% of the time.

Small sample sizes indicate the need to assess the completeness of our dataset and consider what factors affect the likelihood of a forecast being issued. We calculated the percent completeness for each site–year combination, for each site over the entire study period, and for the entire dataset. One-hundred-percent completeness corresponds to a forecast issued every 6 h, 365 days a year. Recall that outside of benchmark locations, the NWS only issues forecasts when increased hydrologic activity is expected. We therefore expect that completeness measures will reflect that a small proportion of possible forecasts were actually issued.

Overall, the dataset is 6.7% complete, but available data are distributed unevenly both between sites and across years. We see from Fig. 6 that sites near the outlet of major rivers and those near population centers are more complete, while basins located upstream have much lower levels of completeness and many have years where no forecasts were recorded. The year-to-year variation in completeness is similar between comparable sites. In particular, most of the basins with large samples show similar patterns over the years; this reflects the pattern of drought and flood years. Although variation in completeness is more pronounced between sites than between years, Fig. 7 illustrates that in some years there is a consistent lack of forecasting across all sites. The most noticeable is 2005, coinciding with a drought in Iowa. The most complete forecast locations (the benchmark sites) should be at least 14.65% complete for each year of record; however, this is not necessarily the case. Climatic factors such as drought may explain some of the missing data. For almost all sites, more forecasts were issued in 2008 than any other year, corresponding to historic flooding. Time series plots of forecasts also show that a large majority of forecasts were issued between spring and early fall, reflecting the seasonality of flooding in Iowa (Villarini et al. 2011a).

Flood forecasts are of most immediate importance when the river is above flood stage; thus, the NWS concentrates prediction efforts when flooding is expected. Figure 8 illustrates that the probability that a forecast was issued is conditional on flood stage. In general, the higher

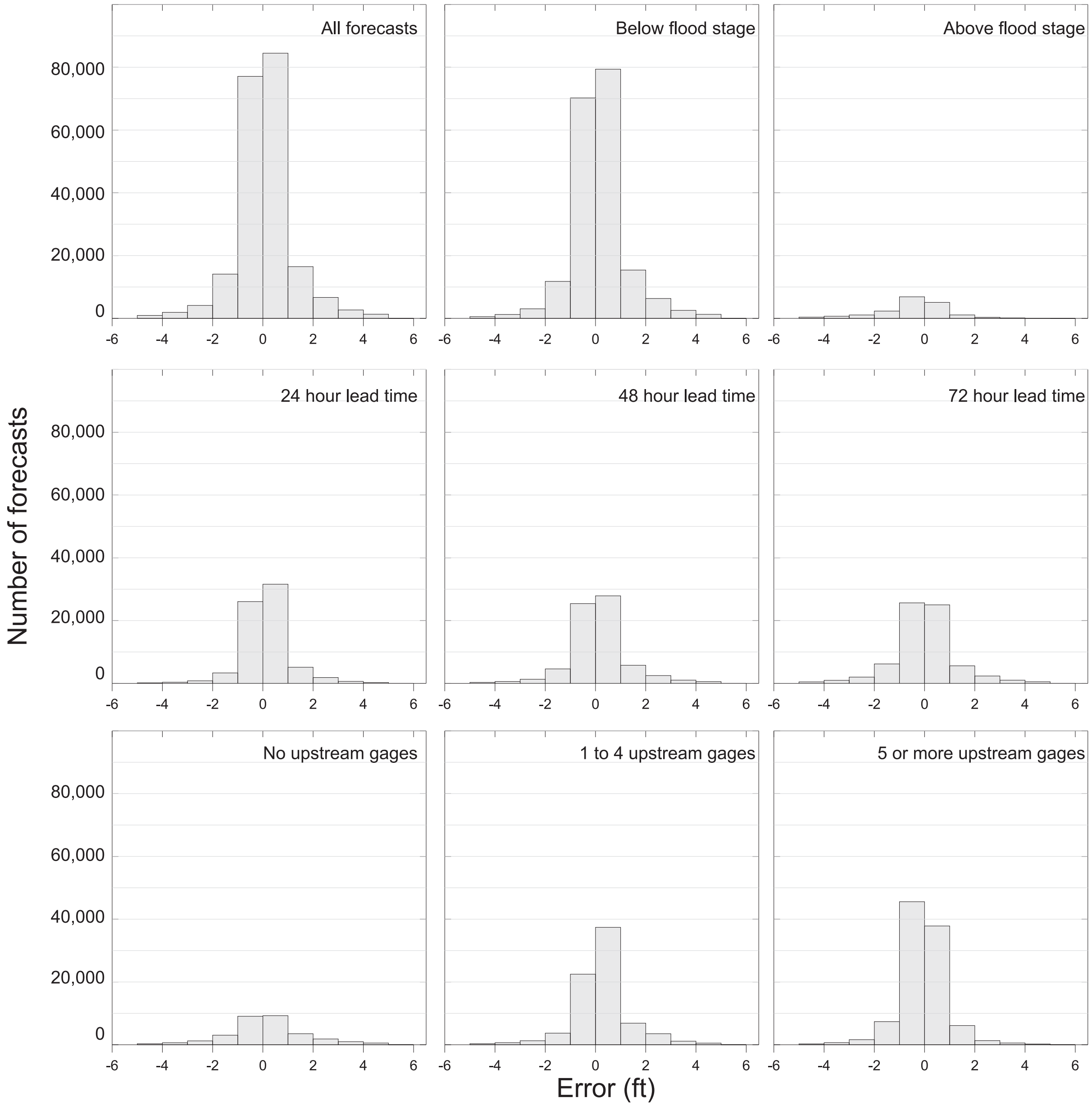

FIG. 11. Distribution of the errors in stage forecasts conditioned to flood category, lead time, and availability of upstream gauges.

the flood stage of an observation, the more likely it is that a forecast corresponding to that observation will exist. In Fig. 8 and the following figures, the colors correspond to the color code utilized by the NWS Advanced Hydrologic Prediction Service for action stage, flood stage, moderate flood stage, and major flood stage. There is not a strong relationship between basin area and the likelihood of forecasts at flood stages. For observations below flood stage, sites with small drainage areas have an extremely low likelihood of forecast. This reflects the practice of the NWS to issue forecasts for smaller basins only when flooding is expected, unless emergency managers request that the NWS issue routine forecasts year round at those sites. Above the threshold of approximately 2000 mi^2 (5180 km^2), there is more variability, as some basins have a higher probability of forecast compared with smaller basins.

Though there is considerable variation, there is some relationship between the completeness of the record and basin size, with larger basins likely to have higher levels of completeness (Fig. 9). In part this reflects the choice of benchmark locations, which are intended to be geographically significant. Most of the gauges with less than 5% completeness have an upstream area of less than

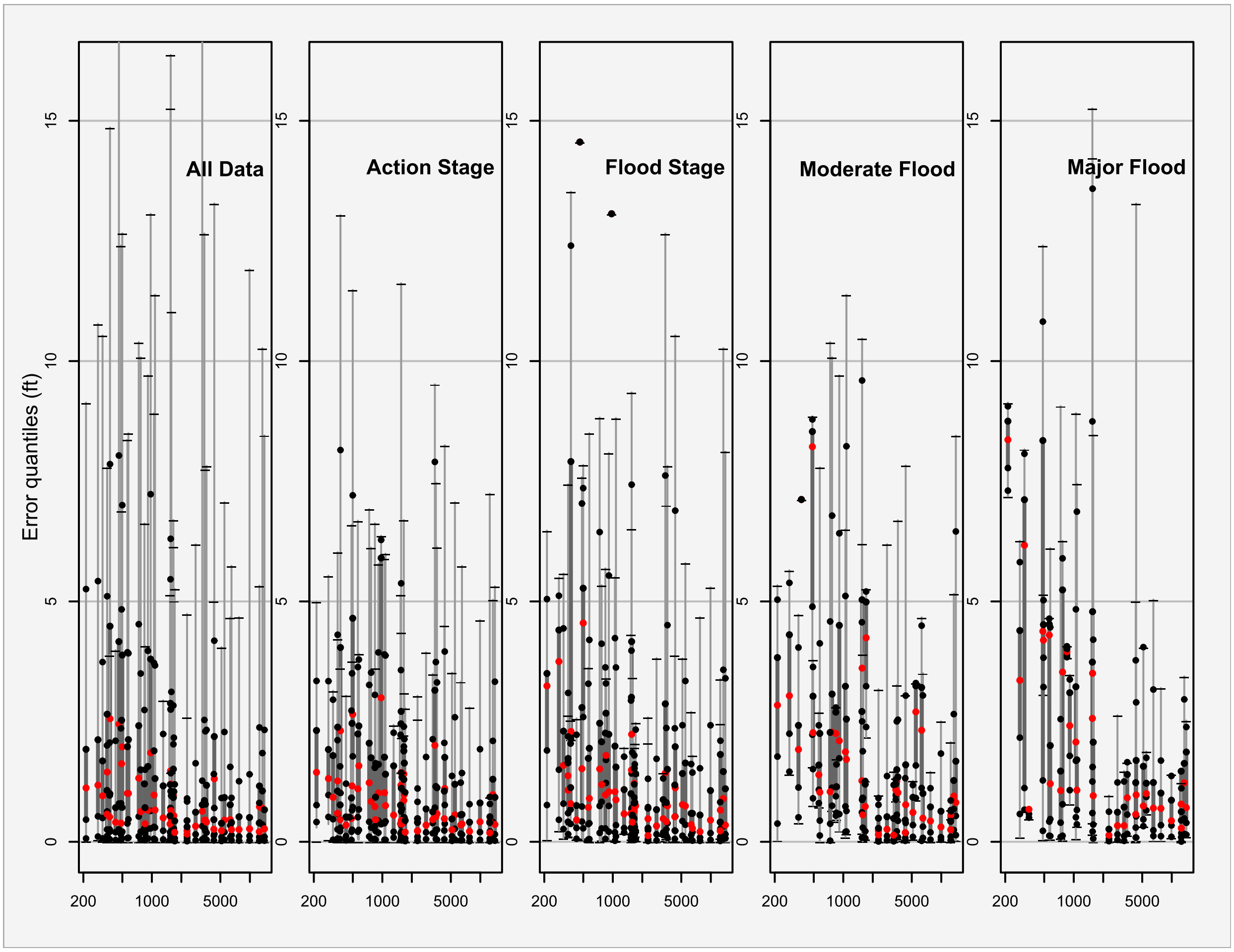

FIG. 12. Significant quantiles of 1-day forecast error data at each flood stage for each site, plotted against the upstream drainage area. Median is marked by the red dot; 75th percentile is the large black dot; and 5th, 25th, and 95th percentiles are small black dots. Minimum and maximum are horizontal bars.

2000 mi^2, and almost no basins of this size have higher completeness. The site TRCI4 is an outlier, at less than 1% complete with an upstream basin area > 10 000 mi^2 (25 900 km^2). Although Tracy (TRCI4) is located along the river's main stem, it has a population of less than 1000.

5. Forecast skill analysis

We performed the same analysis as Welles et al. (2007) for our dataset. They used persistence as a reference forecast to compare unskilled (persistence) forecast errors to skilled (NWS) forecast errors and found that NWS forecasts below flood stage showed skill at 1-, 2-, and 3-day lead times. NWS forecasts above flood stage showed skill at 1-day lead time but the improvement made to the persistence forecast decreased with lead time and was negligible for the 3-day forecast. When we repeated the analysis on the Iowa forecast–observation dataset (Fig. 10), we found that NWS forecasts below flood stage were skillful (as demonstrated by lower RMSEs) compared with persistence forecasts for 1-, 2-, and 3-day forecasts (Table 2). However, above flood stage, where Welles et al. (2007) only saw skill in 1-day forecasts, we found that the improvement of the actual NWS forecasts over persistence forecasts actually increased as lead time increased. In other words, when compared with persistence, the 3-day NWS forecast was more skilled than the 1-day forecasts. Like Welles et al. (2007), we did not see evidence of increasing skill over the study period.

Thus, for our dataset we have (at least partially) affirmatively answered the question of whether the NWS forecasts add skill to the river stage predictions. Now we wish to explore what factors influence forecast skill. We do so by performing various conditional analyses of the forecast error dataset. We expand on previous

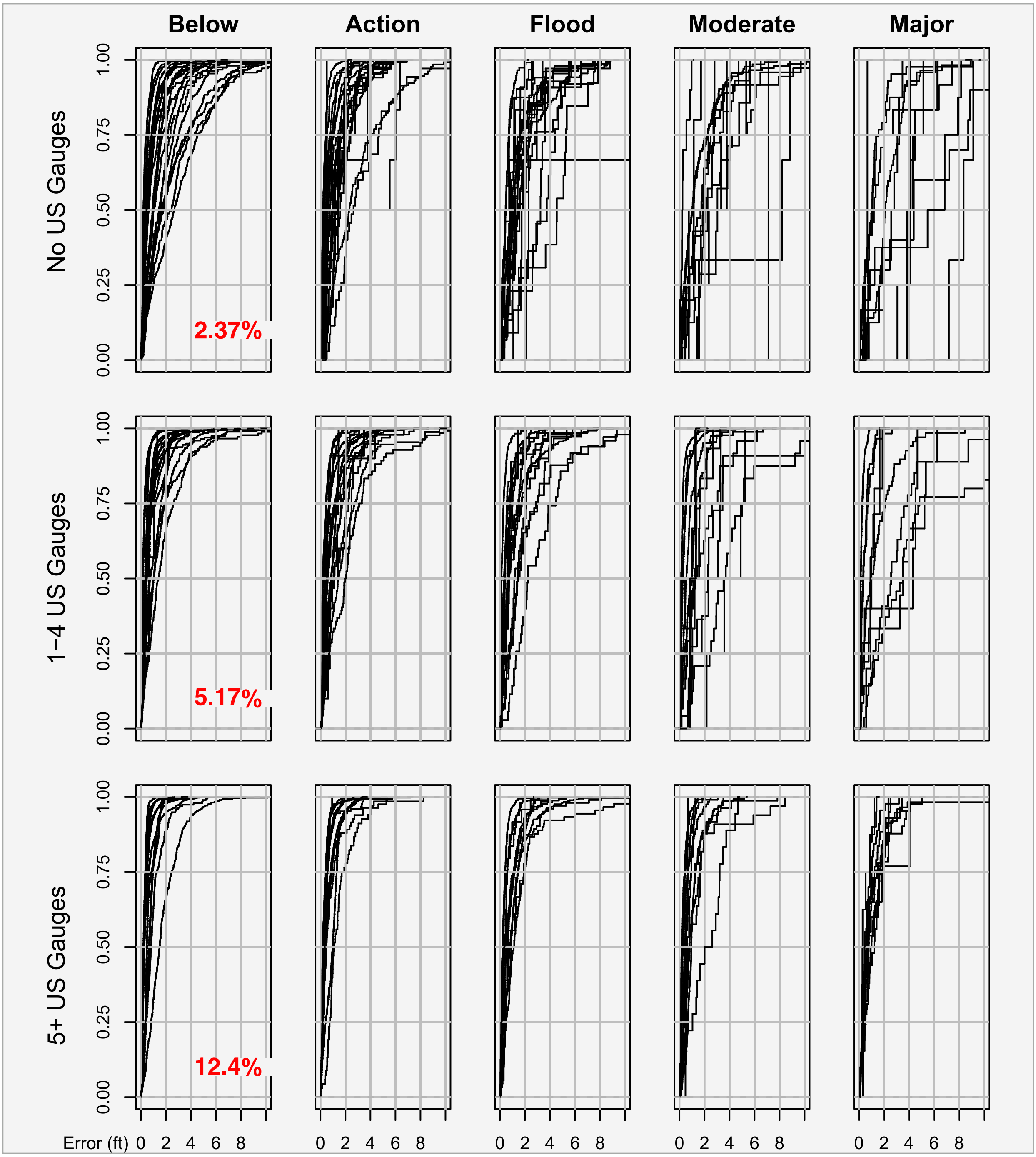

FIG. 13. ECDFs of prediction error from 1-day forecasts at each flood stage, for each site. Sites are divided into three categories by the number of additional gauges upstream. Only gauges at NWS forecast sites used in this study are counted; 23 sites have no upstream gauges, 17 sites have between one and four upstream gauges, and 11 sites have five or more upstream gauges. Average completeness for each subset of sites is in red.

analyses of forecast error by comparing across different basins and under different conditions. We also consider the error distributions rather than single statistics, as shown in Fig. 11.

The distribution of errors below flood stage is fairly smooth and rarely exceeds 5 ft (1.5 m) in range. The error distributions skew right and peak close to zero. Larger sites with more observations tend to be especially

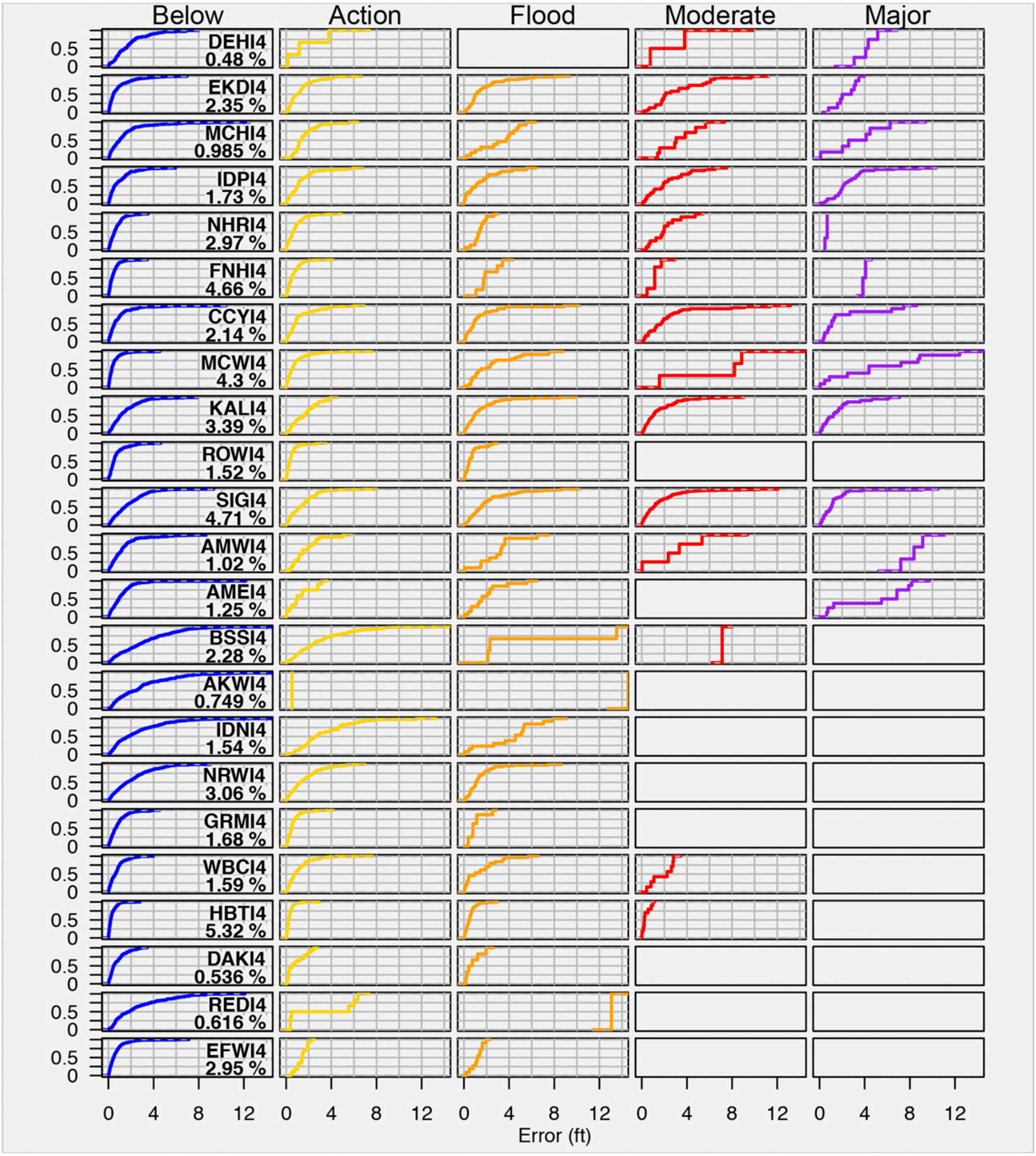

FIG. 14. ECDFs of 1-day prediction errors at each flood stage and overall completeness for each site with no upstream gauges.

strongly peaked. As the flood level changes, so does the distribution of errors. At higher flood stages, the distribution of errors becomes noisy as a result of the much smaller number of observations. Many sites, particularly those with smaller drainage areas, have a few or no observations at the higher flood levels, making the information about the error distribution at these stages difficult to infer. During flood events, we can see that the range of errors and the spread of the distribution tend to increase; that is, forecasts are less accurate. The peak of the distribution also increases slightly. Forecast error distributions corresponding to longer lead times are similarly shaped but, as expected, there is a tendency for the distribution to be less peaked, the range of errors to

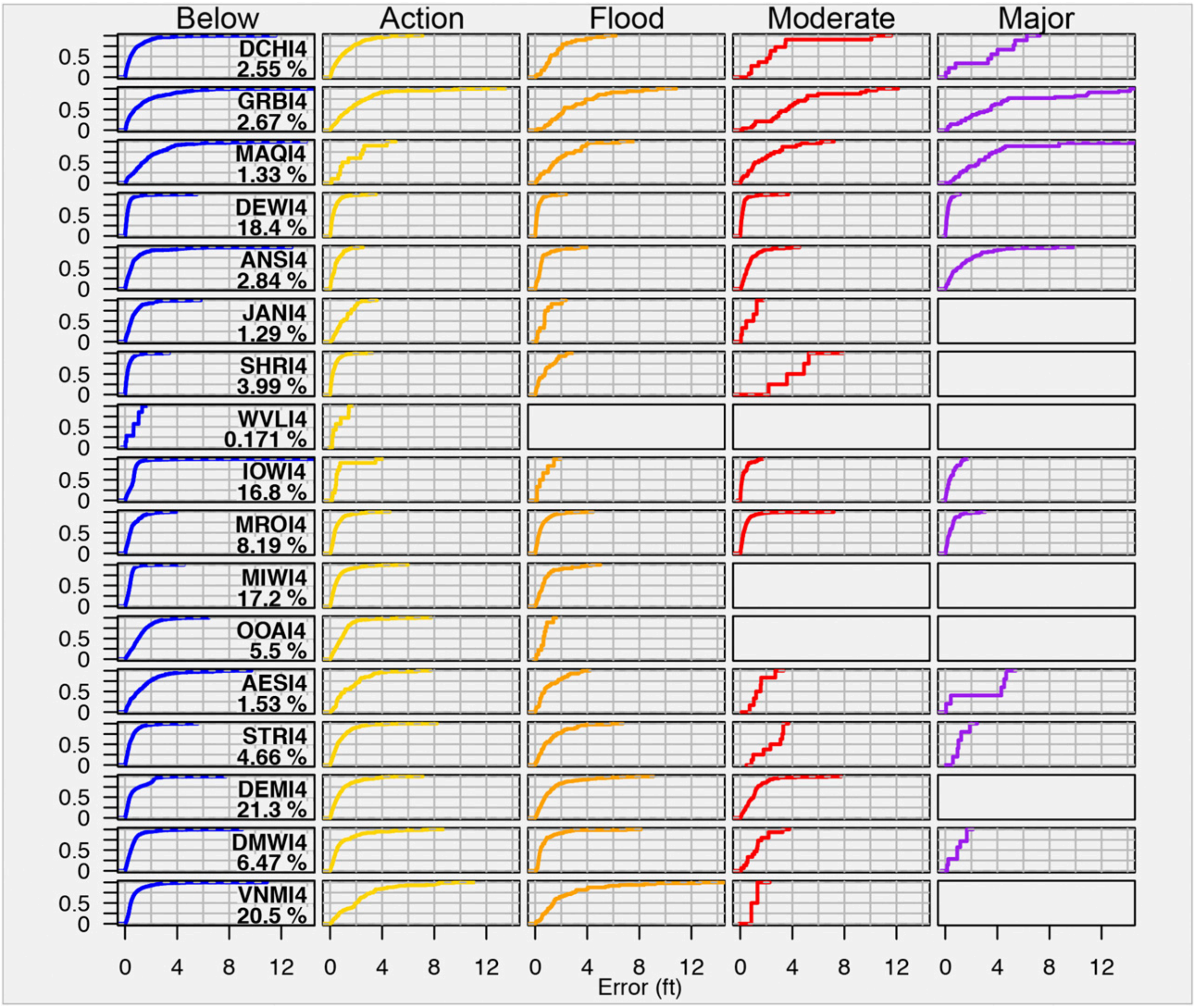

FIG. 15. As in Fig. 14, but for each site with between one and four upstream gauges.

increase, and the center to shift. That is, stage predictions are overall less accurate when there is a greater lag between the forecast issue time and the projected event.

Below flood stage there is generally little difference between the distributions of positive and negative errors. A positive error represents an overprediction (the forecasted stage is higher than the observed stage), and a negative error is an underprediction (forecasted stage is lower than the observed stage). At higher flood levels, underprediction accounts for more large errors. When the lead time is longer, the center and range of the underprediction errors tend to increase compared with the distribution of overprediction errors.

Figure 12 displays quantiles of forecast error distribution for all forecast sites. Lower quantiles for large basins are small, with median errors of less than 1 ft. Smaller basins are more variable; while some are comparable to large sites, others have significantly higher median errors (and higher values of other quantiles), and the disparity increases with flood stage. For larger basins, the median error tends to increase slightly with flood stage, but the range of errors tends to decrease. We next consider in more detail how site characteristics discussed earlier (drainage area, presence of gauges upstream, completeness of record) affect the forecast error distribution via empirical cumulative distribution functions.

Figure 13 groups the error cumulative density function (ECDF) of the forecast error from 1-day forecasts by the number of gauges located upstream. Sites with no upstream gauges have the lowest overall completeness (2.37%) and the greatest variation in skill below flood level. The most skillful sites display accuracy similar to some of the well-gauged sites (i.e., sites with five or more upstream gauges), where the proportion of prediction

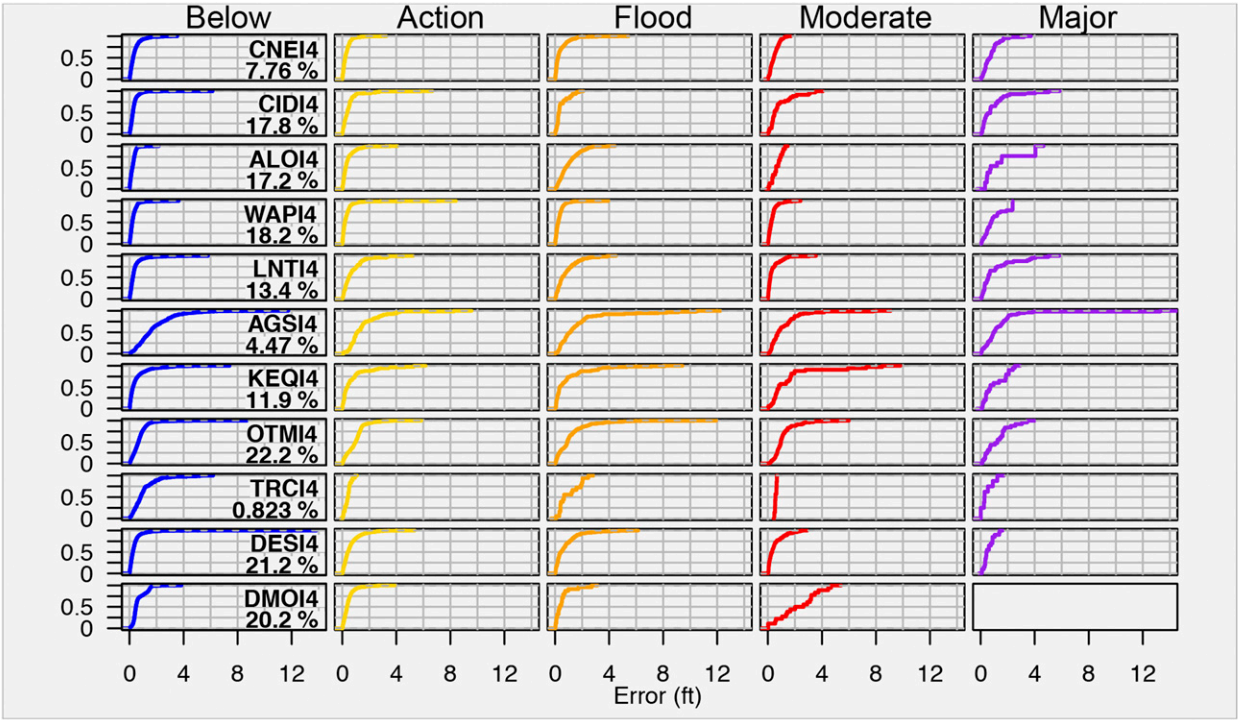

FIG. 16. As in Fig. 14, but for each site with five or more upstream gauges.

errors under 2 ft (0.61 m) approaches 100%, below flood level. The ECDFs clearly display the tendency of forecast errors to increase as flood stage rises, and the discontinuity of the curve reflects the decreased sample size. In western Iowa in particular (sites from the Des Moines, Skunk, and Raccoon River basins), many sites have no observations of events overpassing the moderate or major flood stage (Fig. 14).

The pattern of decreasing skill as categories shift from action to major flood stage is also noticeable for sites with between one and four upstream gauges. Particularly at higher flood stages, a few of the sparsely gauged sites do not experience a notable decrease in forecast accuracy, but most do. For two sites in this category, DEWI4 and IOWI4, the proportion of errors with magnitude less than 2 ft even at major flood stage is close to 100% (Fig. 15). Both are notable for the high completeness of their records, 18.4% and 16.8%, respectively, where the average completeness for sparsely gauged sites is 5.17%. IOWI4 also has the advantage of flow control directly upstream.

Well-gauged sites average 12.4% completeness and behave differently than sparsely gauged sites at higher flood levels. Although skill tends to decrease as flood stage increases, the difference between well-gauged sites below flood level and at flood levels is less substantial. In some cases, (notably CNEI4 and AGSI4, which are the sites located farthest downstream in their respective basins, and DESI4, which is located at the outlet of the Raccoon River as it empties into the Des Moines River), forecast skill does not appear to decrease at higher flood stages (Fig. 16). In general, well-gauged sites have enough observations at high flood stages that the shape of the distribution is visible.

Similarly, Fig. 17 groups ECDFs of forecast error by the size of the forecast point's upstream drainage area. A majority of the sites with no upstream gauges and those with small basin areas [less than 1000 mi^2 (2590 km^2)] overlap. All sites with an upstream area > 10 000 mi^2 (25 900 km^2) have many upstream gauges. As such, the small-basin error distributions resemble the zero-gauge distributions, with forecast skill varying significantly between sites, and skill decreasing as flood stage increases. Forecast datasets from small sites are on average less complete (2.16%) than any other subset of sites; as such, ECDFs with a negligible number of observations at flood levels tend to belong to the small-basin category (Fig. 18). As a group, large basins are among the most skilled, displaying comparatively little reduction of forecast skill at higher flood levels (Fig. 19).

The use of precipitation forecasts, with their high degree of uncertainty, likely affects the forecast accuracy in smaller basins more strongly than in large basins, where geographic precision matters less. This contributes to a

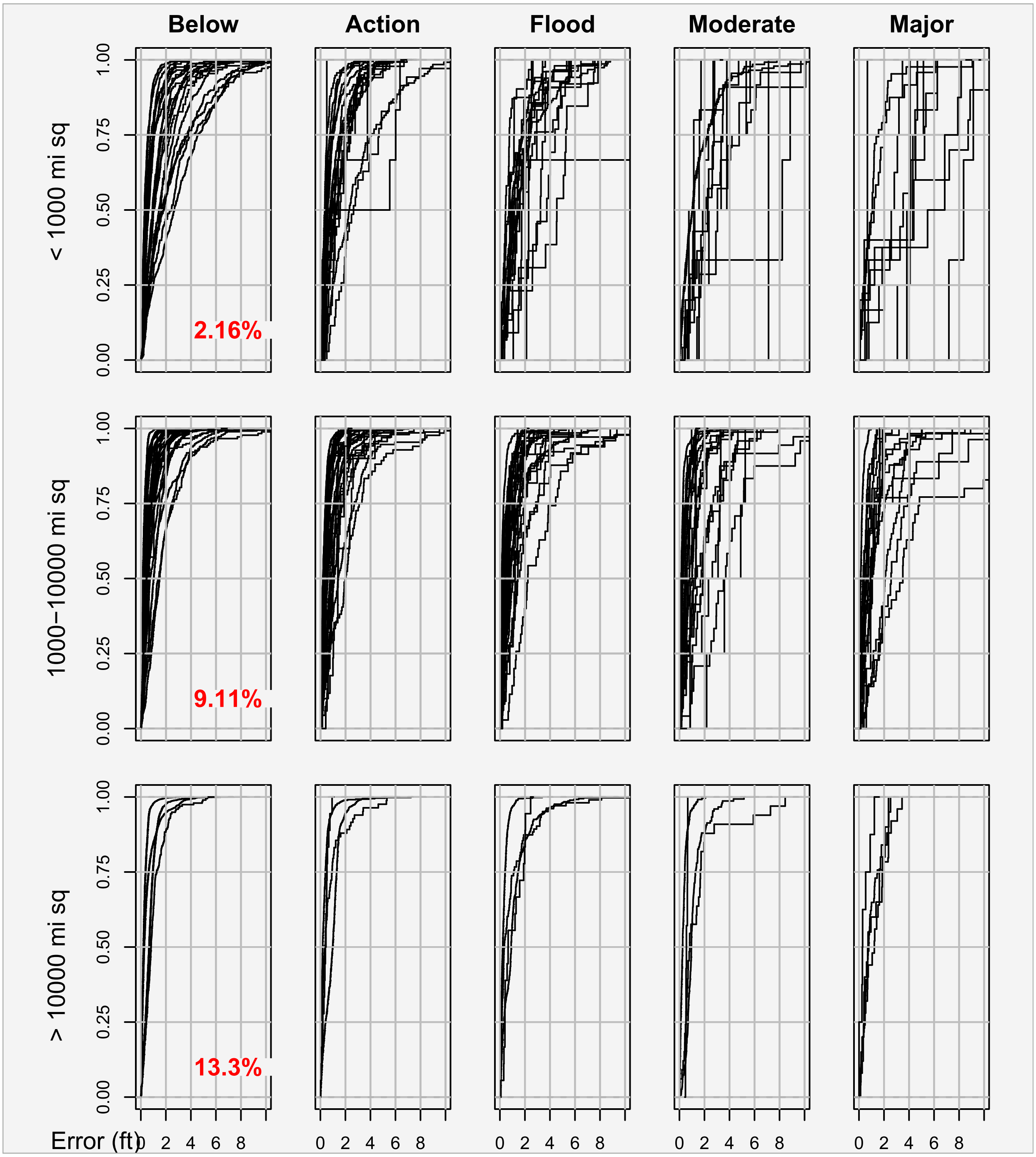

FIG. 17. ECDFs of prediction error from 1-day forecasts at each flood stage, for each site. Sites are divided into three categories by upstream drainage area: 20 sites have upstream area $< 1000\,\mathrm{mi}^2$, 27 sites have area 1000–$10\,000\,\mathrm{mi}^2$, and 4 sites have area $> 10\,000\,\mathrm{mi}^2$. Average completeness for each subset of sites is in red.

greater decline in forecast skill in small basins as the flood stage (and thus the intensity of the hydrologic activity at the time) increases. However, other factors must also affect forecast error, as some small basins are comparable to large ones in forecast skill.

Rainfall effect assessment

We are also interested in knowing how expected forecast errors may vary under different weather conditions. The effect of future rainfall on forecast error is

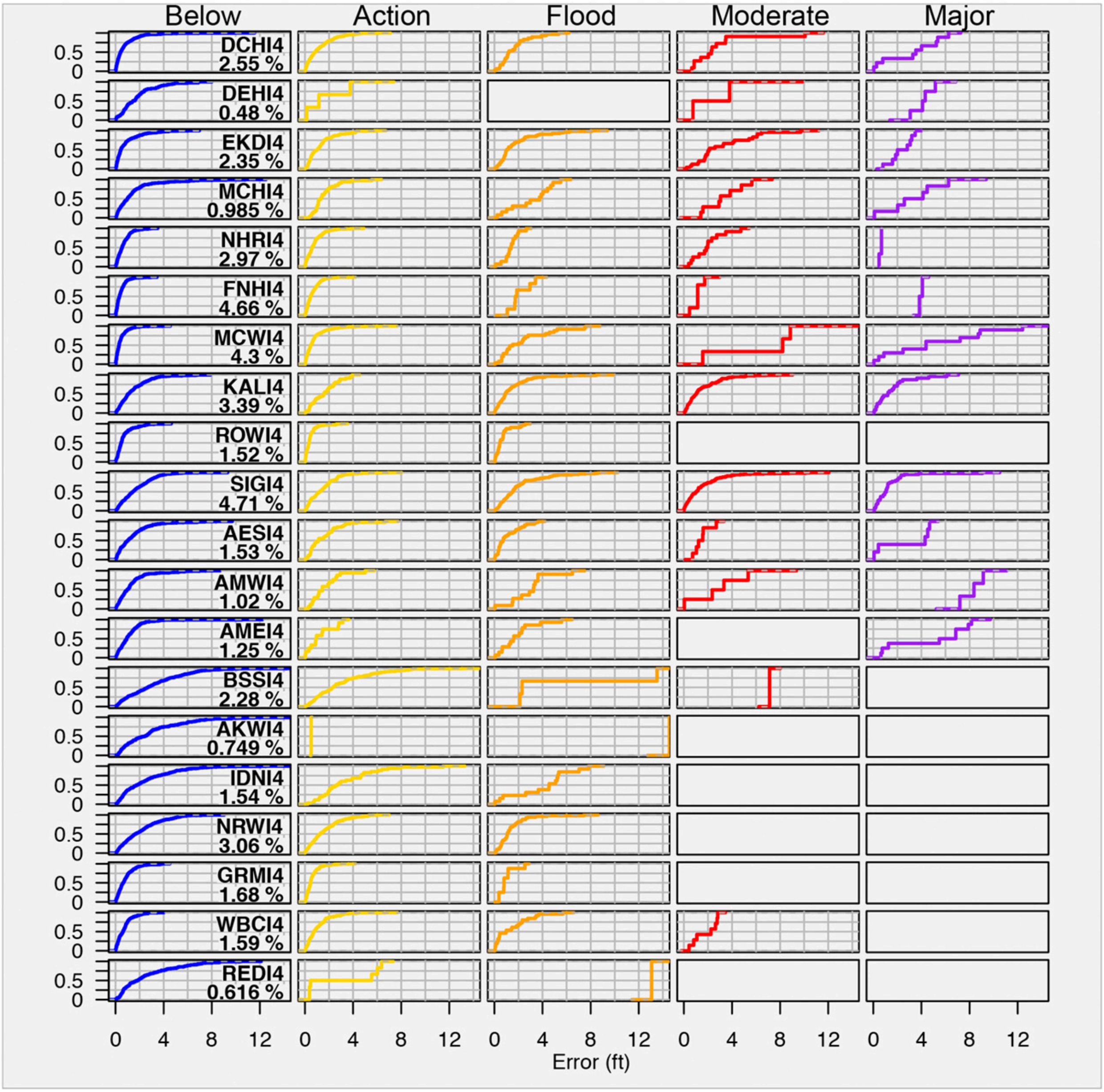

FIG. 18. ECDFs of 1-day prediction errors at each flood stage, for each site with basin area < 1000 mi^2. Overall completeness for each site is given.

of particular concern as a result of the possibility of flooding. Because of the difficulty in obtaining precise, accurate weather forecasts, we hypothesize that rainfall that occurs after a stage forecast is issued increases the error in that forecast. Efforts to forecast increases in the river stage must take into account rainfall of unknown quantity, timing, and location, and thus are more difficult than predicting the falling limb of a hydrograph. To investigate the effect of rainfall, we calculate the RMSE for the set of forecasts where rainfall occurred within a specified length of time (time window) after forecast issue time and compare with the RMSEs for cases where no rain fell in the window after issue time. Rain that falls after a forecast is issued may or may not have been expected, but its exact characteristics and thus the impact on the river stage are fundamentally unknown. We examined how RMSE evolves with respect to the length of the rainfall time window, the forecast lead time, and the amount of rainfall that occurred in the window and we considered the relationships to basin characteristics.

Figure 20 shows how RMSE varies with an increasing time window. By itself, the choice of time

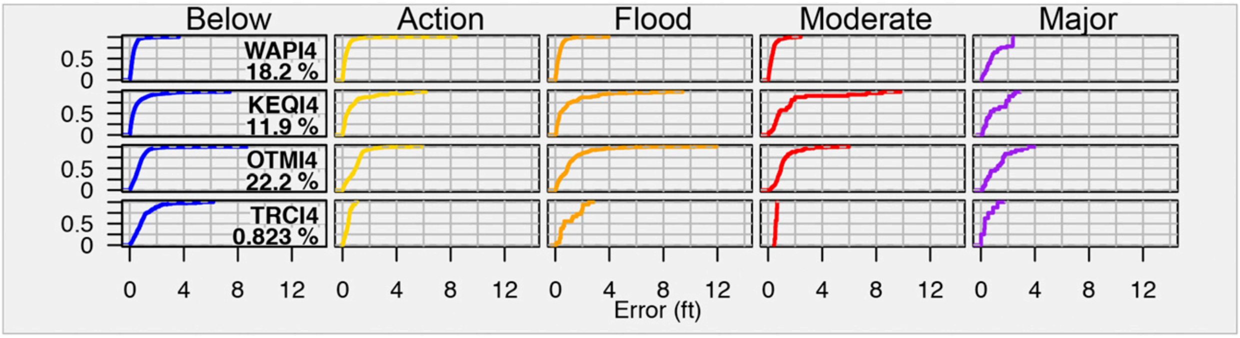

FIG. 19. As in Fig. 18, but for each site with basin area > 10 000 mi^2.

window for rainfall does not have a clear effect on RMSE. For almost half the sites at the 48-h lead time, there is no significant impact; the RMSEs with and without rainfall are nearly equal. Where a difference is observed, we see that, as expected, the RMSEs of the forecasts followed by rainfall tend to be higher than those of forecasts not followed by rain. Increasing the lead time of the forecast and the rainfall threshold (the amount of rainfall required for a forecast to be considered in the set of forecasts followed by rain) tends to increase the distance between RMSEs with and without rainfall but does not produce more additional insight.

There are multiple timing factors that affect the choice of time window and inform our interpretation. Runoff from anywhere in the drainage basin will take time to reach the river. From that point, it could take as long as the basin's maximum travel time to reach the gauge; however, for a localized storm in a larger drainage basin, the actual travel time could vary significantly. The choice of rainfall window has different implications for different sites. For instance, using a small rainfall window for a large basin with a long travel time may mean that rainfall that remains in the basin long enough to affect the river stage is not considered. It is also important to consider the relationship between the rainfall window and the lead time of the forecast. If the lead time is longer than the rainfall window, the stage at the forecasted time (time $t + \tau$, where t is the issue time and τ is the lead time) may be largely unaffected by the rainfall that we are considering, because the water has already passed through the basin. If the rainfall window is longer than the lead time, we are considering rainfall accumulation that is no longer relevant, since the forecasted time has already passed.

We examined the effect of scaling the rainfall window by defining it as the maximum travel time for each site. Figure 21 shows the RMSEs with and without rainfall in the travel time window for each site, plotted against lead times from 6 h to 5 days. We see that RMSEs with and without rainfall are both small and nearly equal when the lead time is less than a day; however, at longer lead times the magnitude of the difference between RMSE with rainfall and RMSE without does not have a consistent relationship with basin size. In some cases, the difference increases steadily with lead time, in some the difference remains relatively consistent or appears to level off, and in some cases no pattern is discernable. For two sites (AKWI4 and INDI4), the RMSE without rainfall exceeds the RMSE with rain. Such cases may be explained by the data limitations. AKWI4 is a small site with small sample size. Splitting the data by rainfall further reduces the number of observations with which we can calculate the RMSE (especially for the forecasts without rainfall, which are fewer in number), so the high value of the RMSE for one lead time may not be representative.

Geomorphological properties of the drainage network may also make accurate forecasting more difficult. For instance, the basin of the Wapsipinicon River (see Fig. 3) may illustrate the importance of considering basin shape. This basin has a strongly atypical shape, being long and narrow, which affects the travel time and rate at which water moves through the basin to its outlet. Such basins may not be well described by traditional routing models.

We expect that for small basins, in particular, rainfall makes river stage difficult to predict since there is little or no advance warning from rainfall upstream, and the stream itself will respond quickly to rainfall. The rainfall threshold is the minimum amount of rainfall within the window that must occur for the forecast to be included in the set of forecasts with rain. Figures 22 and 23 plot the RMSEs with rainfall against thresholds of 0, 0.25, 0.5, 1.0, 1.5, and 2.0 in. (with lead times of 72 and 120 h, respectively). We see clearly that overall, the

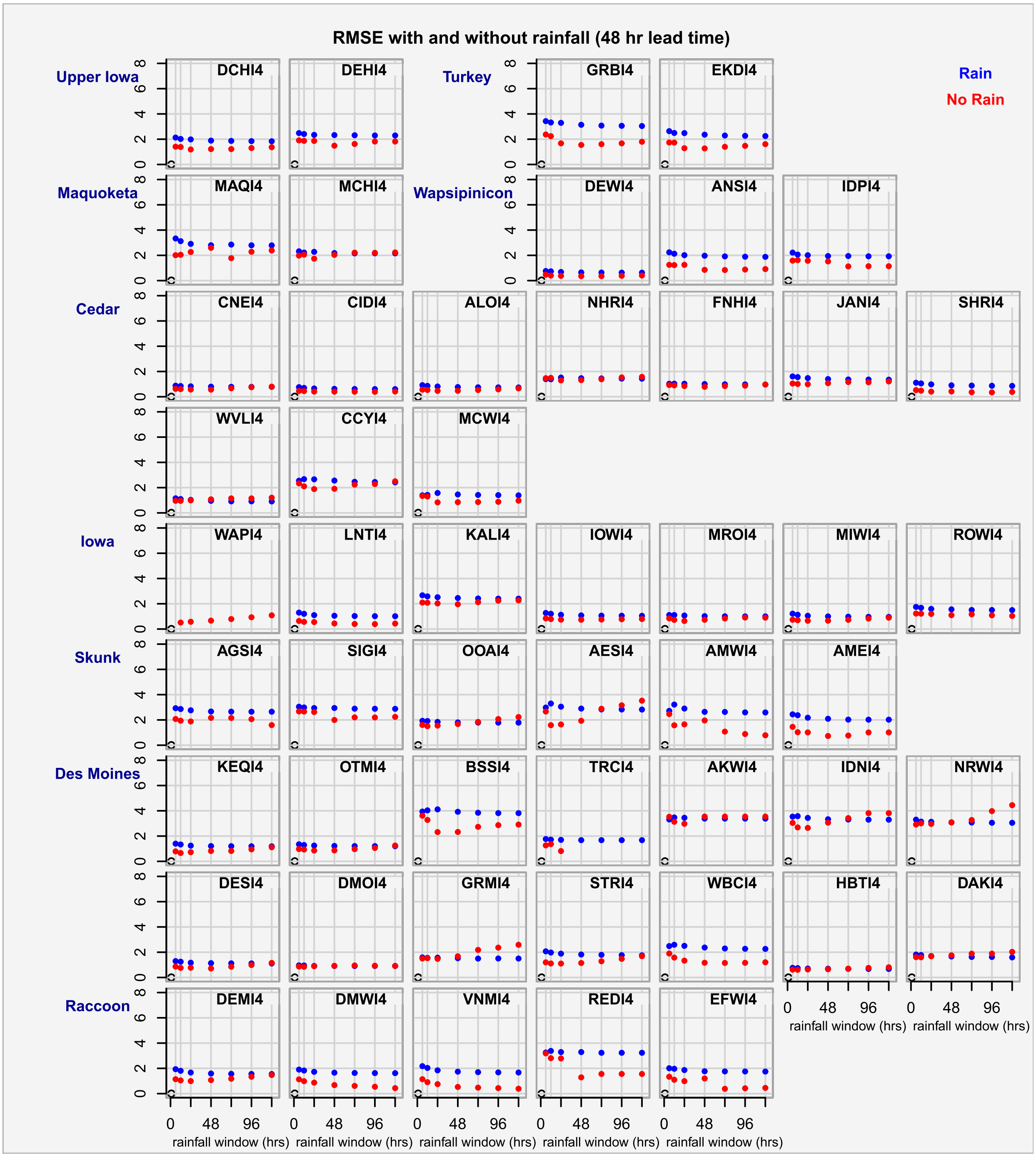

FIG. 20. RMSE of 48-h (2 day) forecasts at each site. Points in red represent RMSE calculated for forecasts in cases where no rain fell during the specified time window after the forecast issue time. Points in blue represent RMSE calculated where there was rainfall within the specified time window after the forecast was issued. The rainfall threshold was 0 in. and any amount of rainfall placed a forecast in the "rain" category. RMSEs are plotted against time windows of 6, 12, 24, 48, 72, 96, and 120 h.

magnitude of the difference between RMSEs with rain and RMSEs without rain increases with rainfall threshold; thus, there is a relationship between the RMSE and the quantity of rainfall.

We also consider whether having a lead time that is less than or greater than the basin travel time changes the impact of rainfall on RMSE. We compare Figs. 22 and 23 for basins where the 72-h lead time is less than the

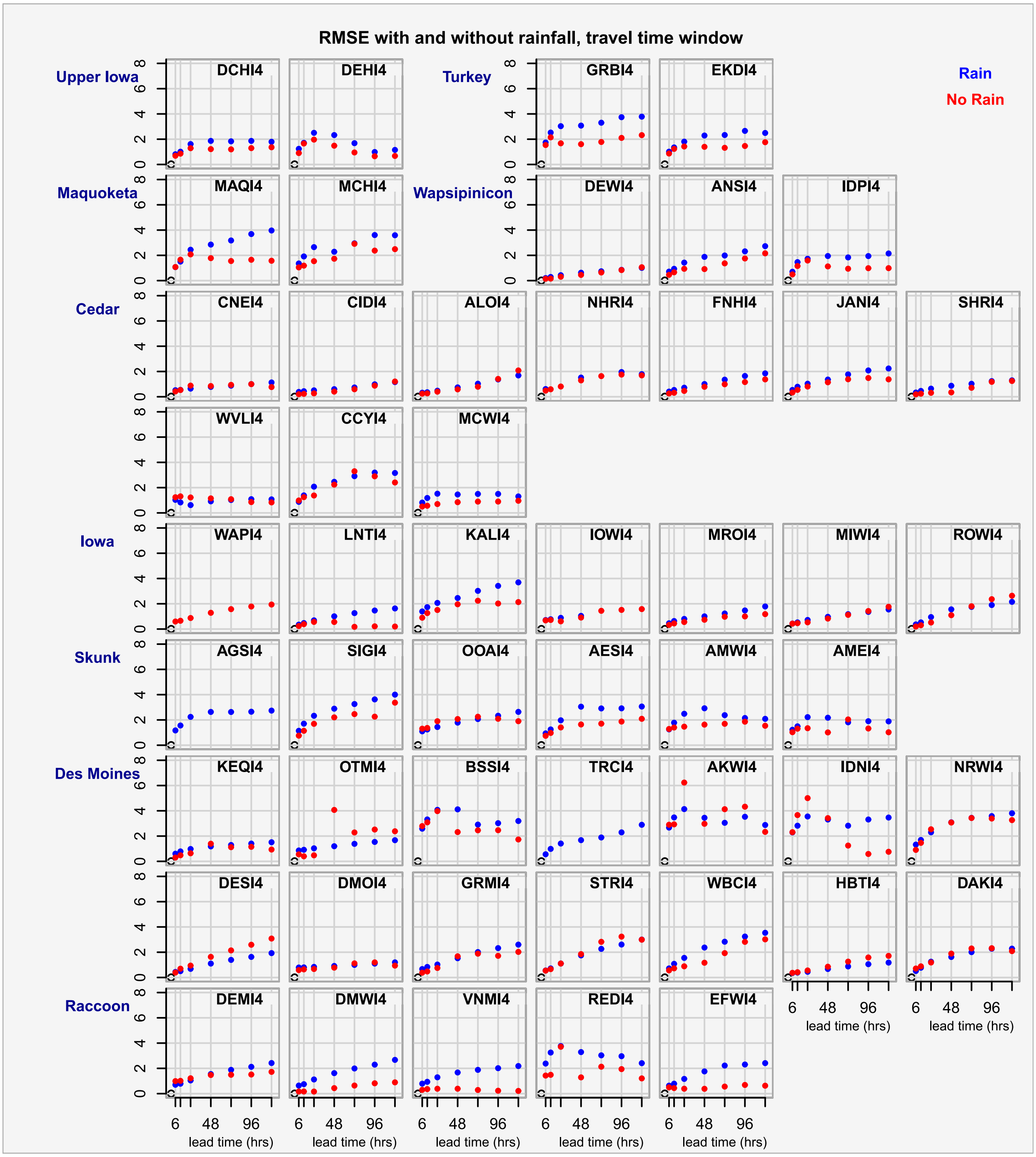

FIG. 21. RMSE of forecasts at each site. Points in red represent RMSE calculated for forecasts in cases where no rain fell during the travel time window following the forecast issue time. Points in blue represent RMSE calculated where there was rainfall within the travel time window after the forecast was issued. The rainfall threshold was 0 in. RMSEs are plotted against lead times of 6, 12, 24, 48, 72, 96, and 120 h.

travel time and the 120-h lead time is greater than the travel time. For these basins, RMSEs with rainfall are generally closer to the no-rain RMSE and are less dependent on the rainfall threshold, though the difference may not be significant. This implies that the accuracy of stage predictions with rainfall at a particular lead time may depend on the relationship between lead time and basin size. However, because of the

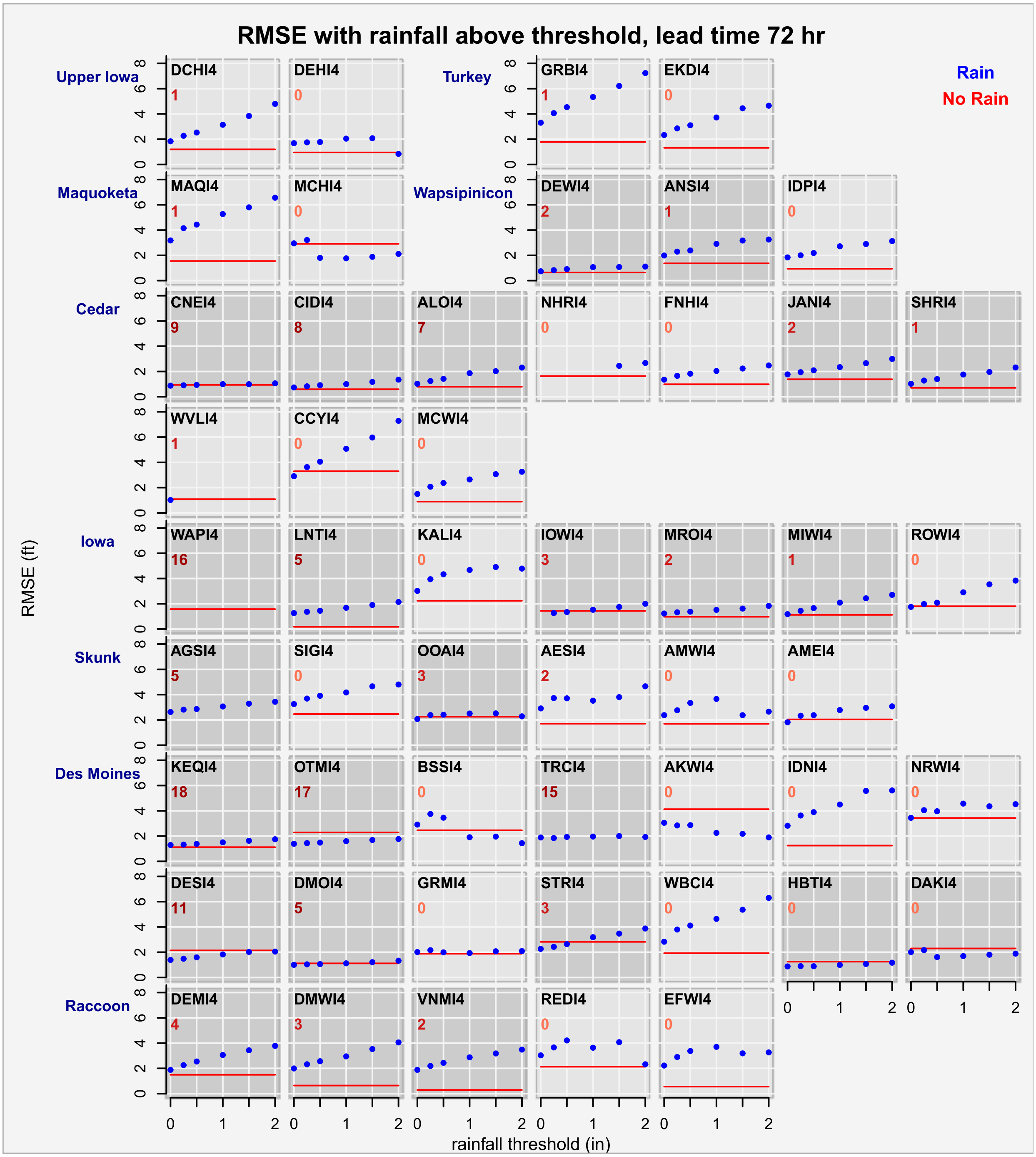

FIG. 22. RMSE of 72-h forecasts at each site. Red line is the RMSE for forecasts in cases where no rain fell during the travel time window following the forecast issue time. Points in blue represent RMSE calculated where there was a minimum quantity of rainfall (horizontal axis) within the travel time window after the forecast was issued. Location ID and the number of gauges upstream of each location are shown in the top-left corner of each plot. Light gray background indicates that maximum travel time for the site is less than (or equal to) the forecast lead time; darker gray background indicates that maximum travel time is greater than the lead time.

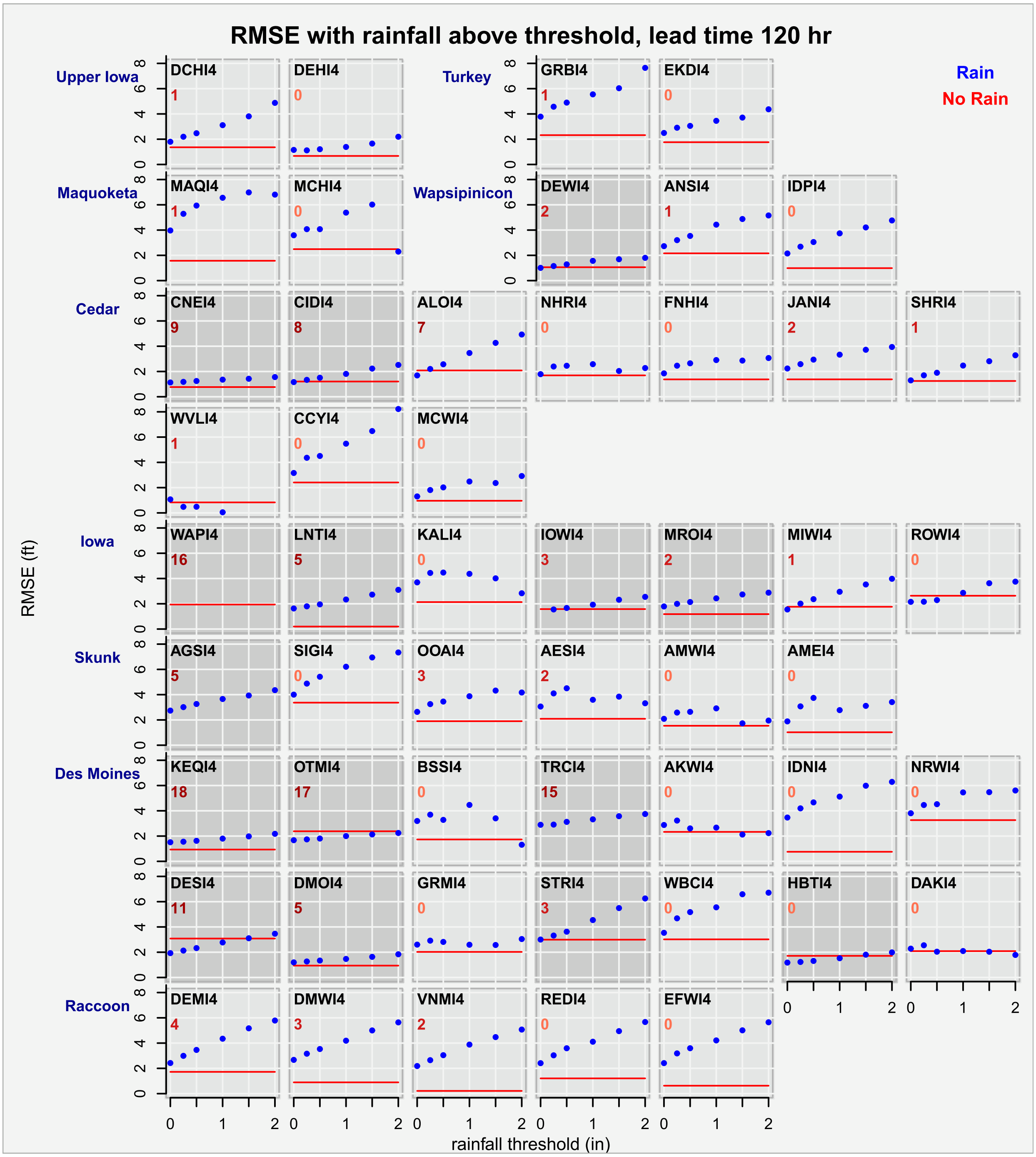

FIG. 23. As in Fig. 22, but for RMSE of 120-h forecasts at each site.

spatial variability of rainfall, the basin response time is variable as well.

6. Conclusions

Strong predictive relationships between physical characteristics of a basin (travel time, upstream drainage area), quantities (rainfall after forecast issuance), or factors related to forecasting efforts (number of upstream gauges, completeness of record) and forecast skill have not been identified. However, certain patterns are observed. Smaller/poorly gauged sites exhibit greater variation in forecast skill, but are not necessarily less skillful than large/well-gauged sites below flood level. At high flood levels, the decrease in skill for smaller/poorly gauged sites is more pronounced, but there are still a few sites that have skill equivalent to the largest/well-gauged sites. Factors not considered in this study, such as experience forecasting at a particular location, characteristics of poorly forecasted events, the influence of upstream controls, or geophysical characteristics of a particular basin, may influence forecast skill.

We found agreement with Welles et al. (2007) on the behavior of forecast errors below flood stage, but in contrast to that study, we found actual forecasts to be better than persistence for predictions above flood stage, except for the 24-h lead-time forecasts where the skill is similar. While this study considers only a relatively small region, it adds considerably to the scant literature documenting operational flood forecasting uncertainty in the United States. Given recent efforts at the National Water Center toward modernizing the practice of forecasting using automated models, having benchmarks such as the ones provided herein will be vital for measuring the future improvement of forecast skill.

Acknowledgments. The authors are grateful to staff of the NWS North Central River Forecast Center for making the forecast data available. GZ and FQ were supported by the Iowa Flood Center. WK was supported by the Rose and Joseph Summers endowment.

REFERENCES

Anderson, E. A., 1976: A point energy and mass balance model of a snow cover. NOAA Tech. Rep. NWS 19, Silver Spring, MD, 150 pp.

Brunner, G., 2010: HEC-RAS river analysis system: Hydraulic reference manual, version 4.1. Hydrologic Engineering Center Rep. CPD-69, U.S. Army Corps of Engineers, Davis, CA, 411 pp. [Available online at http://www.hec.usace.army.mil/software/hec-ras/documentation/HEC-RAS_4.1_Reference_Manual.pdf.]

Burnash, R. J. C., 1995: The NWS river forecast system: Catchment modeling. *Computer Models of Watershed Hydrology*, V. P. Singah, Ed., Water Resources Publications, 311–366.

Klipsch, J. D., and M. B. Hurst, 2007: HEC-ResSim reservoir system simulation user's manual version 3.0. Hydrologic Engineering Center Rep. CPD-82, U.S. Army Corps of Engineers, Davis, CA, 512 pp.

Krajewski, W. F., and Coauthors, 2016: Real-time flood forecasting and information system for the state of Iowa. *Bull. Amer. Meteor. Soc.*, **98**, 539–554, https://doi.org/10.1175/BAMS-D-15-00243.1.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.]

Mallakpour, I., and G. Villarini, 2015: The changing nature of flooding across the central United States. *Nat. Climate Change*, **5**, 250–254, doi:10.1038/nclimate2516.

NOAA, 2016: National Water Model: Improving NOAA's water prediction services. NWS Office of Water Prediction, 2 pp. [Available online at http://water.noaa.gov/documents/wrn-national-water-model.pdf.]

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, doi:10.1175/WAF-D-13-00066.1.

Villarini, G., J. A. Smith, M. L. Baeck, and W. F. Krajewski, 2011a: Examining flood frequency distributions in the Midwest U.S. *J. Amer. Water Resour. Assoc.*, **47**, 447–463, doi:10.1111/j.1752-1688.2011.00540.x.

——, ——, ——, R. Vitolo, D. B. Stephenson, and W. F. Krajewski, 2011b: On the frequency of heavy rainfall for the Midwest of the United States. *J. Hydrol.*, **400**, 103–120, doi:10.1016/j.jhydrol.2011.01.027.

Welles, E., and S. Sorooshian, 2009: Scientific verification of deterministic river stage forecasts. *J. Hydrometeor.*, **10**, 507–520, doi:10.1175/2008JHM1022.1.

——, ——, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 503, doi:10.1175/BAMS-88-4-503.